

Deep Learning for Infrared-Based Photovoltaic Module Fault Classification in Drone Inspection: A Critical Review of Datasets, Models and Evaluation Practice, with a Binary-Classification Case Study

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

Drone-borne infrared (IR) thermography combined with deep learning is widely used for automated photovoltaic (PV) inspection. However, reported accuracies of approximately 85–94% on public benchmarks provide limited evidence of reliability under field conditions. This article critically reviews IR-based PV fault classification and detection around three questions: what is predicted (task granularity), from what data (resolution, provenance, and class structure), and how success is measured (evaluation protocol). We synthesize six previous surveys, five data resources, and recent research on convolutional networks, transformers, self-supervised learning, lightweight models, object detectors, multimodal language models, and federated learning. Three findings emerge. First, reported accuracy on crop-level benchmarks often falls within a relatively narrow range despite substantial differences in model size, while evidence from an industry-scale study highlights dataset diversity as a major determinant of performance. Second, common evaluation practices— random crop-level splits, class merging, balanced prevalence, and single-site data—limit the conclusions that benchmark results can support. Third, the translation of sensitivity and specificity into maintenance workload is rarely quantified. To illustrate this issue, we reanalyse a binary module-level case study comprising 1,523 image crops. ResNet18 achieves 89.6% accuracy, compared with 76.1% for a small convolutional neural network. Assuming that

sensitivity and specificity remain unchanged, reducing fault prevalence to 2% yields an estimated precision of approximately 16% for ResNet18, corresponding to roughly five false alarms per correctly detected faulty module. We conclude with a reporting checklist and a research agenda for more reliable evaluation and deployment.

Keywords: photovoltaic fault detection; infrared thermography; UAV inspection; deep learning; benchmark datasets; evaluation protocols; generalization.

Introduction

Photovoltaic generation depends on routine inspection because faults such as hot spots, bypass-diode failures, open-circuit substrings and string outages reduce yield and can create fire risk. Outdoor infrared thermography of modules and plants in operation is standardized in IEC TS 62446-3, which specifies measurement equipment, ambient conditions, inspection procedure, reporting, personnel qualification and a matrix of thermal abnormalities [1]. Small unmanned aerial systems and crewed aircraft have made plant- scale acquisition practical; one industry study reports airborne infrared inspections of 6,155 North American solar farms [2]. Deep learning is the natural tool for turning such imagery into maintenance decisions, and the literature on it has grown quickly.

Several surveys already exist. Early work established the physics of IR diagnosis of crystalline-silicon modules [3]; later reviews addressed aerial IR inspection [4], gigawatt-scale imaging [5], imagery pattern recognition [6], and panel overlay and fault detection [7]. A 2026 review spans electrical, visible, IR, electroluminescence, UAV and multimodal sensing and stresses that results from different studies are not directly comparable because datasets, fault taxonomies and validation protocols differ [8]. What remains under-examined is the chain that links task granularity, data properties and evaluation protocol for module-level IR classification, which is the form in which most public benchmarks are cast, and the consequences of that chain for deployment.

This article makes four contributions:

- 1. A task hierarchy (L0–L4) that places datasets, models and deployed systems on a common map (Section 4);

- 2. A verified comparison of public IR datasets and of headline results reported on them, with explicit caveats (Sections 5–7);
- 3. An analysis of evaluation threats and a reporting checklist (Section 8); and
- 4. A quantitative case study showing how sensitivity and specificity measured on a balanced test set translate into precision and false-alarm workload at other prevalences (Section 9), followed by a research agenda (Section 10).

Scope and Review Method

This is a critical narrative review, not a PRISMA systematic review, and no results are pooled. Literature was identified in September–October 2026 through searches of publisher sites, preprint servers and dataset repositories, combining terms for PV, infrared or thermal imaging, UAV or aerial inspection, fault or defect classification or detection, and deep learning, supplemented by backward citation chasing from recent surveys and benchmark papers. Priority went to (i) dataset papers and dataset records, (ii) benchmark or comparative studies from 2023–2026, (iii) large-scale field-deployment studies and (iv) prior surveys. Electrical-only diagnosis (I–V curves, SCADA) and electroluminescence work were considered only for contrast. Three limitations apply. The search is not exhaustive and may under-represent non-English and industrial reports. Several recent sources are preprints that have not been peer reviewed; these are marked as preprints in the reference list. Where a result is known to us only through another paper’s summary, the text says so (for example, results on the Infrared Solar Modules benchmark summarized in [2]). Reported numbers are quoted as reported, with their task and protocol where these were available.

Physical Basis, Acquisition and Fault Taxonomy

What infrared imaging can and cannot see

Thermography observes the thermal consequence of a fault, not its cause. Current mismatch, open circuits and bypass-diode activation redistribute dissipation within a module and appear as localized or patterned heating. Because different mechanisms can produce similar patterns, and because apparent temperature depends on irradiance, ambient temperature, wind and viewing geometry, thermal-image interpretation is sensitive to acquisition conditions [8]. IEC TS 62446-3 therefore

defines passive outdoor inspection under natural sunlight, distinguishes simplified from detailed inspections, and is commonly cited as requiring a minimum irradiance of 600 W/m² [1]. Data collected outside such conditions may carry different class-conditional appearance, which matters for any learned model.

Operational pipelines add their own conventions. In the at-scale system of [2], a hot spot is a localized anomaly more than 5 °C above neighboring cells, and severity tiers run from below 5 °C to 15 °C or more; the authors also note that time of day, season, weather and soil temperature make a single global temperature normalization infeasible and use a multi-step normalization pipeline instead.

Label spaces are not interoperable

Public datasets partition faults differently. The Infrared Solar Modules (ISM) benchmark has 12 classes (one no-anomaly and eleven anomaly classes) [9]; PVF-10 has ten categories [10]; the sUAS record used in our case study has eight (seven fault types plus normal) [11]; and the at-scale pipeline detects six defect types, from hot spots and diode bypass to string outages and tracker misalignment [2]. Some categories describe appearance (hot spot, multi-cell), others an inferred electrical cause (open circuit, short circuit). Results therefore cannot be compared across datasets without an explicit label mapping, and binary “normal versus fault” tasks derived from different taxonomies are not the same task.

Spatial resolution constrains diagnostic information

The diagnostic information available in thermal imagery depends on spatial resolution, module dimensions, and the extent of the thermal anomaly. The ISM benchmark contains module crops of 24 × 40 pixels [9], whereas PVF-10 reports a spatial resolution of approximately 0.02 m/pixel [10]. At this resolution, a module measuring 1.2 × 1.7 m would occupy approximately 60 × 85 pixels, assuming a near-normal viewing angle and uniform ground sampling distance. The sUAS dataset was annotated using an orthomosaic with a ground sampling distance of 3.13 cm/pixel [11], corresponding to approximately 54 × 32 pixels for a 1.7 × 1.0 m module under the same geometric assumptions. The industry-scale study reports an infrared ground sampling distance of up to 6.5 cm/pixel [2]. These geometric

estimates should be distinguished from stored crop dimensions, which may be affected by perspective correction, resampling, and preprocessing.

Spatial resolution is particularly consequential for anomalies whose thermal signatures occupy only a small fraction of a module. Resizing a low-resolution crop to a larger network input does not recover spatial detail absent from the original acquisition. In a comparative CNN–transformer benchmark on ISM, electrical fault categories achieved F1 scores above 0.90, while soiling achieved approximately 0.20–0.33; the study attributes part of this disparity to limitations in thermal imaging resolution [12]. This finding suggests that architecture selection alone may be insufficient when the relevant diagnostic signal is weak or poorly resolved. However, resolution should not be treated as the sole explanation: class imbalance, annotation uncertainty, acquisition conditions, and overlap between fault appearances may also influence performance. Consequently, benchmark comparisons should report both the original image resolution and the physical scale of the anomalies being classified.

Task Formulations: A Five-Level Hierarchy

The term “fault detection” encompasses several distinct prediction tasks, each requiring different inputs, annotations, and evaluation criteria. To clarify these distinctions, we organize the inspection workflow into five levels:

- L0: Data acquisition. Collection of thermal frames, videos, or orthomosaics, together with acquisition metadata and image-quality assessment.
- L1: Spatial localization. Identification of PV tables, individual modules, or anomalous regions through object detection or segmentation.
- L2: Binary condition classification. Assignment of an individual module crop to a normal or faulty category according to the dataset’s labeling criteria.
- L3: Fault-type classification. Identification of a specific anomaly or fault category within a defined taxonomy, allowing multiple labels where faults coexist.

- L4: Operational assessment and decision support. Assessment of fault severity, estimation of associated energy losses, and prioritization of maintenance actions using thermal evidence and relevant operational information.

This hierarchy describes functional stages rather than a mandatory sequence: some systems combine localization and classification, while others operate directly on pre-extracted module crops. Nevertheless, performance at one level does not establish effectiveness at another. Accurate crop classification at L2 or L3 does not demonstrate reliable localization in full aerial images, correct identification of an electrical cause, or useful maintenance prioritization at L4.

Most public crop-based benchmarks primarily evaluate L2–L3. Their operational relevance therefore depends on additional evidence concerning acquisition quality, localization errors, generalization to unseen sites, and the consequences of false alarms and missed faults. Evaluation should identify the level addressed explicitly and distinguish component-level predictive performance from the effectiveness of the complete inspection workflow.

Crop-level classification (L2–L3) offers a comparatively straightforward setting for annotation and benchmarking, but it has two structural limitations. First, it assumes that an upstream process has already localized and extracted the relevant modules, so classification performance alone does not capture errors introduced during localization. Second, isolated crops restrict the spatial context available for diagnosis. In operational inspection, the temperature of a potentially anomalous region may be assessed relative to neighboring modules; this comparison cannot be reproduced from an isolated module crop without additional contextual information [2].

Detection-based systems operating at L1, or jointly at L1 and L3, can preserve surrounding scene information and provide spatially explicit predictions, although they generally require more detailed annotations. The industry-scale pipeline described in [2] combines table segmentation, oriented module detection, and anomaly detection before translating the outputs into an operational-efficiency rating. It therefore illustrates how several stages of the L0–L4 hierarchy can be integrated within a single inspection workflow. Earlier studies on orthorectified UAV imagery [13] and

module extraction and rectification from aerial IR video [14] addressed complementary requirements for connecting image acquisition with automated module assessment.

Data Resources

Provenance varies. ISM and PVF-10 aggregate crops from many modules, and PVF-10 spans eight plants, making variation across acquisition conditions explicit [10]. The sUAS record, by contrast, is annotated on a single orthomosaic of one array [11], while the Thessaloniki dataset originates from one farm [15]. Single-array datasets are valuable for controlled experiments, but evaluation confined to one array cannot establish generalization to unseen sites. Dataset diversity should therefore be assessed through acquisition provenance and site coverage, rather than image count alone.

Class distributions are imbalanced. In the sUAS record, the normal class contains 2,835 images, approximately 3.4 times the combined fault classes (840 images), while individual fault categories contain between 2 and 431 images. Binary merging, as used in several studies, offers a pragmatic response to limited samples: a derived healthy-versus-faulty task combines nine PVF-10 fault categories into one class [16], and a binary ISM task contains exactly 10,000 samples per class [12]. However, merging changes the prediction target and conceals differences in performance among fault types. It does not, by itself, change the overall proportion of faulty modules when all samples are retained; changes in binary prevalence arise from sampling, filtering, or balancing. These distinctions are consequential for evaluation and deployment (Sections 8 and 9).

Strong baseline performance warrants scrutiny. The authors of the industry-scale study report that a random-forest baseline achieves 84.0% overall accuracy for anomaly classification on ISM and suggest that the benchmark may be insufficiently challenging for assessing deployable models [2]. This interpretation should be distinguished from direct evidence of field performance: a strong baseline result alone does not establish that a dataset is unsuitable. Nevertheless, it emphasizes the importance of comparing complex models with simple baselines under identical evaluation conditions. Such comparisons help determine whether

additional architectural complexity yields meaningful improvements and whether those improvements persist under more demanding validation protocols.

Model Families and Reported Results

Convolutional networks and lightweight designs

Custom and off-the-shelf CNNs are widely used for crop-level classification. Reported results include 92.5% accuracy for a custom CNN on ISM [17], 93% for EfficientNet-B0 features combined with a support-vector machine on a 20,000-crop dataset [18], 88% for a cascading ensemble with a voting CNN [19], and 85% for a NASNet–LSTM combination [20]. These results are cited here as summarized in [2] and should not be interpreted as directly comparable without examining their label definitions, data partitions, and evaluation protocols.

Research on lightweight models includes a coupled U-Net/DenseNet design intended to reduce model complexity while maintaining PV fault-classification performance [21]. A pruned, low-rank tiny residual network designed for Edge TPU deployment reports 94.73% accuracy for anomaly detection and up to 88.44% for multiclass diagnosis across benchmark configurations containing 2, 8, 11, and 12 classes. The study compares its approach with 59 existing CNN models [22]. Earlier research also investigated deep networks for remote anomaly detection and classification [23]. Claims of deployment efficiency should nevertheless distinguish reductions in parameter count from measured latency, memory consumption, and energy use on the target hardware.

Residual networks and transfer learning

ResNet-family backbones [24] are common baselines and feature extractors. Applications include module classification after extraction from IR videos [14] and table segmentation within the industry-scale inspection pipeline [2]. Their use as pretrained backbones also supports transfer learning when labeled thermal imagery is limited, although the benefit depends on the source representation, adaptation strategy, and target dataset.

A comparative study evaluates ResNet-18, EfficientNet-B0, ViT-Tiny, and Swin-Tiny on the approximately 20,000-image ISM dataset under binary and multiclass formulations [12]. The study additionally uses XRAI saliency analysis to examine

whether influential image regions correspond to expected thermal signatures. Reported patterns include localized hot spots for cell defects, linear thermal patterns for diode failures, and thermal boundaries associated with vegetation shading [12]. Such analyses can help assess whether predictions are consistent with plausible diagnostic evidence. However, saliency agreement alone does not establish causal reasoning or exclude reliance on spurious features; it should complement quantitative evaluation and controlled robustness tests.

Transformers and self-supervised learning

Vision transformers [25] have been evaluated alongside CNNs for PV fault classification [12]. The industry-scale study reports that transformer-based alternatives did not provide substantial improvements over its CNN components while increasing training complexity and computational cost [2]. This observation is specific to the evaluated pipeline and should not be generalized to all transformer architectures or acquisition settings.

A more recent approach uses self-supervised pretraining. JEFFNet combines a JEPA- pretrained vision transformer with EfficientNetV2-S features and is evaluated on PVF- 10 and ISM for multiclass classification and derived binary tasks [16]. On PVF-10, the study reports 93.86% accuracy and an F1 score of 93.69%, with 108.6 million parameters, approximately 47.3% fewer than the 205.91 million parameters reported for the GEPFNet comparator. The averaging convention used for F1 should be specified when comparing these results with other studies.

Self-supervised learning offers a means of exploiting unlabeled thermal imagery when verified fault annotations are scarce. Its practical value, however, requires evaluation on independent sites and acquisition conditions, together with disclosure of the pretraining data and any overlap with downstream evaluation samples.

Detectors and segmentation

For plant-scale inspection, object detectors provide spatially explicit predictions by localizing modules or anomalous regions directly in aerial imagery. The YOLO family [26] underpins the anomaly detector described in [2], whose largest variant reports a metric denoted as precision@50 of 89.8%. This notation should be defined

according to the original study and distinguished from average precision at an intersection-over-union threshold of 0.50.

A staged RGB–IR pipeline segments arrays in visible images and classifies defects using a ResNet. As summarized in [8], it reports 91.7% mAP@0.5, 95% defect-detection accuracy, and a processing rate of 36 frames per second on 3,000 paired images from a 65 MW plant. These metrics describe different aspects of performance and should be interpreted alongside the corresponding task definitions and evaluation procedures.

A comparative study evaluates YOLOv10, YOLOv12, RT-DETR, RF-DETR, and the open-vocabulary models YOLO-World and OWL-ViT on a public thermal UAV dataset [27]. RF-DETR achieves the highest reported mAP@0.50 of 82.6% at 12.6 frames per second, whereas YOLO-World achieves 78.1% at 31.3 frames per second. This corresponds to a difference of 4.5 percentage points in mAP and an approximately 2.5-fold difference in throughput. The comparison illustrates a deployment trade-off between localization performance and processing speed, although throughput remains specific to the hardware, input resolution, batch size, and timing procedure used in the study.

Multimodal language models

A comparison of GPT-4o and ResNet on aerial IR imagery considers binary and multi-class classification tasks [28]. The study reports advantages for ResNet in computational efficiency and implementation simplicity, while GPT-4o provides greater flexibility and textual explanations of subtle anomalies. These findings suggest a possible complementary workflow in which a compact classifier performs high-volume screening and a multimodal language model supports explanation and triage.

However, fluent explanations should not be treated as evidence of diagnostic correctness. Evaluation should assess classification performance and explanation reliability separately, including whether generated descriptions are grounded in observable thermal features. Reproducible comparisons also require disclosure of the model version, prompts, image preprocessing, inference settings, and computational or service costs.

Privacy-preserving and federated learning

Restricted access to inspection imagery and incomplete fault coverage at individual sites motivate collaborative learning across PV operators. Federated learning [29] enables participating sites to exchange model updates while retaining their raw data locally. However, this arrangement does not by itself guarantee privacy; protection depends on the threat model and any additional safeguards applied to the exchanged updates.

Within the literature reviewed here, federated approaches remain limited and largely address tasks outside IR module classification. Examples include a decentralized approach to inverter IGBT fault diagnosis motivated by limited, non-identically distributed data and privacy concerns [30], and a federated contrastive self-supervised pretraining scheme for PV defect detection that reports improvements over independent local training [31].

We did not identify a benchmark within the reviewed literature that evaluates federated IR module classification across independent sites. This constitutes a research opportunity (Section 10), particularly for assessing performance under differences in acquisition conditions, fault prevalence, and label availability. Such evaluations should compare federated training with local and centralized baselines and report predictive performance, communication costs, and the privacy protections actually implemented.

Synthesis: What the Evidence Supports

Finding 1: Similar headline accuracies occur across substantially different model designs. Reported deep-model accuracies on the crop-level benchmarks summarized in Table 3 fall approximately between 85% and 95%, despite substantial variation in architecture and model size. Examples range from a custom CNN [17] to a hybrid transformer–CNN model containing 108.6 million parameters [16]. A random-forest baseline achieves 84.0% accuracy, numerically one percentage point below the 85% reported for a NASNet–LSTM model [2, 20]. However, differences in task definitions and evaluation protocols prevent these values from establishing either an accuracy plateau or architectural equivalence. The observed range is consistent with the hypothesis that acquisition quality, spatial resolution, and annotation uncertainty

constrain performance. The low reported F1 scores for soiling [12] provide supporting evidence for investigating this hypothesis, but controlled experiments are needed to distinguish these effects from architectural and training differences.

Finding 2: Dataset diversity is a major consideration for inspection at scale. The industry- scale study identifies geographically diverse training data spanning climates and ground temperatures as a principal determinant of airborne IR inspection performance [2]. This conclusion highlights the importance of representing operational variation during model development. Nevertheless, it originates from one industrial study and requires independent replication before it can support a general ranking of dataset diversity against architectural choices. Together with the limited cross-site validation identified in the broader review [8], it supports prioritizing evaluation on unseen plants and acquisition conditions.

Finding 3: High within-dataset accuracy provides limited evidence of broader generalization. Within-dataset accuracies approaching 99.9% have been reported for IR fault detection under restricted acquisition conditions, as summarized in [8] from [32], whereas some benchmarks spanning multiple plants report accuracies near 93%. These differences cannot be attributed directly to model quality because the studies differ in data composition, task difficulty, and evaluation procedures. Restricted variation may contribute to higher within-dataset scores, but this remains an interpretation rather than a demonstrated explanation. Consequently, headline accuracy should be assessed alongside acquisition provenance, class distributions, split design, and independent validation.

Finding 4: Efficiency and explanation are increasingly evaluated alongside predictive performance. Research on compact Edge TPU models [22], detector performance–latency trade-offs [27], saliency analysis against expected thermal signatures [12], and multimodal language-model explanations [28] broadens evaluation beyond classification accuracy. These directions address practical requirements for computational efficiency and decision support. However, compactness does not establish suitability for a particular device, and plausible explanations do not establish diagnostic reliability. Stronger deployment evidence

requires measurements on target hardware, validation across independent sites, prospective inspection studies, and assessment of maintenance outcomes.

Conclusion

This study critically reviewed UAV-based infrared inspection and computer vision methods for photovoltaic (PV) fault diagnosis, considering six surveys, five data resources, and a supporting binary classification experiment. The review organized existing approaches into five levels, ranging from image acquisition and module localization to fault detection, fault-type classification, and severity assessment linked to energy loss and maintenance. This framework highlights that detecting a thermal anomaly is an intermediate step: its operational value depends on reliable localization, diagnostic interpretation, and validation against electrical performance.

The supporting experiment demonstrated the potential of deep learning for distinguishing normal and faulty PV modules. On a test set containing 124 faulty and 106 normal images, ResNet achieved an accuracy of 89.57%, detecting 110 faulty modules while producing 14 false negatives and 10 false positives. The smaller CNN achieved 76.09% accuracy, with 108 true positives, 16 false negatives, and 39 false positives. Although their fault sensitivities were comparable, ResNet provided substantially higher specificity, reducing false alarms within the evaluated sample. These findings support its suitability for further investigation as an inspection screening model, while remaining insufficient to establish performance across different installations or acquisition conditions.

The analysis also demonstrated why aggregate accuracy alone cannot establish practical reliability. If fault prevalence were 2%, maintaining ResNet's observed sensitivity and specificity would yield an estimated positive predictive value of approximately 16%, equivalent to about five false alarms per correctly detected fault. This projection assumes that performance remains stable under deployment conditions and therefore represents an illustrative scenario rather than a measured field result. Nevertheless, it emphasizes the importance of prevalence-aware evaluation, threshold selection, and reporting precision, sensitivity, specificity, and uncertainty alongside accuracy.

Several limitations constrain the conclusions. The experimental evaluation used a selected binary subset and a single test split, limiting evidence for generalization across sites, seasons, sensors, and flight configurations. Binary labels also conceal differences among fault mechanisms and severities. Moreover, infrared appearance alone does not establish the electrical cause of an anomaly or quantify its contribution to energy loss. Comparisons among reviewed studies remain affected by differences in datasets, annotation practices, evaluation protocols, and metric definitions.

Future research should prioritize independent site-level testing, repeated evaluations, calibrated predictions, and explicit analysis of rare fault categories. Integrating thermal imagery with RGB observations, environmental measurements, and electrical monitoring could strengthen diagnostic interpretation and support severity estimation. Ultimately, progress toward automated PV energy management requires validated links between visual findings, electrical losses, and maintenance outcomes. UAV-based computer vision offers a promising foundation for this objective, but the present evidence primarily supports automated inspection and fault screening.

References

- [1] IEC TS 62446-3:2017, Photovoltaic (PV) Systems—Requirements for Testing, Documentation and Maintenance—Part 3: Photovoltaic Modules and Plants—Outdoor Infrared Thermography. Geneva: IEC, 2017.
- [2] I. Corley, C. Wallace, S. Agrawal, B. Putrah, and J. Lvowski, “Aerial infrared health monitoring of solar photovoltaic farms at scale,” arXiv:2503.02128, 2025 (preprint).
- [3] J. A. Tsanakas, L. Ha, and C. Buerhop, “Faults and infrared thermographic diagnosis in operating c-Si photovoltaic modules: A review of research and future challenges,” *Renewable and Sustainable Energy Reviews*, vol. 62, pp. 695–709, 2016. doi: 10.1016/j.rser.2016.04.079.
- [4] A. K. V. de Oliveira, M. Aghaei, and R. Rütther, “Automatic inspection of photovoltaic power plants using aerial infrared thermography: A review,” *Energies*, vol. 15, no. 6, Art. no. 2055, 2022.

- [5] I. M. Peters et al., "Infrared imaging of photovoltaic modules: A review of the state of the art and future challenges facing gigawatt-scale PV power stations," *Progress in Energy*, vol. 4, Art. no. 042010, 2022.
- [6] Y. Zefri, I. Sebari, H. Hajji, G. Aniba, and S. El Bouchini-Idrissi, "Applied imagery pattern recognition for photovoltaic modules' inspection: A review on methods, challenges and future development," *Sustainable Energy Technologies and Assessments*, vol. 52, Art. no. 102071, 2022.
- [7] C. Yang, F. Sun, Y. Zou, Z. Lv, L. Xue, C. Jiang, S. Liu, B. Zhao, and H. Cui, "A survey of photovoltaic panel overlay and fault detection methods," *Energies*, vol. 17, no. 4, Art. no. 837, 2024.
- [8] W. Wang, Q. Su, L. Hou, et al., "Intelligent fault detection in photovoltaic systems: Sensing modalities, AI methods, generalization and future trends," *Frontiers in Earth Science*, vol. 14, Art. no. 1949465, 2026. doi: 10.3389/feart.2026.1949465.
- [9] M. Millendorf, E. Obropta, and N. Vadhavkar, "Infrared solar module dataset for anomaly detection," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. Dataset available at: <https://github.com/RaptorMaps/InfraredSolarModules>.
- [10] 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.
- [11] C. Schaefer, "sUAS infrared image dataset for photovoltaic module fault classification," *Zenodo*, version 1, Jun. 2026. doi: 10.5281/zenodo.20860870.
- [12] S. Aksoy, "Benchmarking vision transformers and CNNs for thermal photovoltaic fault detection with explainable AI validation," *arXiv:2509.07039*, 2025 (preprint).
- [13] Y. Zefri et al., "Developing a deep learning-based layer-3 solution for thermal infrared large-scale photovoltaic module inspection from orthorectified big UAV imagery data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, Art. no. 102652, 2022.

- [14] L. Bommès, T. Pickel, C. Buerhop-Lutz, J. Hauch, C. Brabec, and I. M. Peters, "Computer vision tool for detection, mapping and fault classification of PV modules in aerial IR videos," arXiv:2106.07314, 2021 (preprint).
- [15] "Thermal PV panel detection and fault detection dataset for UAV-based inspection," Zenodo dataset record. doi: 10.5281/zenodo.16420123.
- [16] S. Azimidokht, M. Monemi, A. Kharbouch, F. Hamzehaghdam, M. Rasti, J. Aghaei, and E. Kurvinen, "Joint-embedding predictive architecture for solar PV panel fault classification," arXiv:2607.09205, 2026 (preprint).
- [17] R. H. F. 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.
- [18] Z. B. Duranay, "Fault detection in solar energy systems: A deep learning approach," *Electronics*, vol. 12, no. 21, Art. no. 4397, 2023.
- [19] Z. Barraç, I. Sebari, N. Lamrini, K. Ait El Kadi, and I. Ait Abdelmoula, "A cascading decision system for enhanced anomaly classification of large-scale photovoltaic systems using drone's thermal data with class-imbalance problem," *Results in Engineering*, vol. 25, Art. no. 103876, 2025.
- [20] S. Gopalakrishnan, N. I. Abdul Wahab, V. Veerasamy, H. Hizam, and R. A. Farade, "NASNet-LSTM based deep learning classifier for anomaly detection in solar photovoltaic modules," *Journal of Physics: Conference Series*, Art. no. 012006, IOP Publishing, 2024.
- [21] R. F. Pamungkas, I. B. K. Y. Utama, and Y. M. Jang, "A novel approach for efficient solar panel fault classification using coupled UDenseNet," *Sensors*, vol. 23, no. 10, Art. no. 4918, 2023. doi: 10.3390/s23104918.
- [22] R. M. Reddy, A. Pallikonda, and R. D. A. Raj, "Pruned and low-rank optimized tiny residual architecture for solar PV fault classification on Edge TPU," *ASME Journal of Energy Resources Technology*, 2025. doi: 10.1115/1.4069350.
- [23] M. Le, D. K. Nguyen, V.-D. Dao, N. H. Vu, H. H. T. Vu, et al., "Remote anomaly detection and classification of solar photovoltaic modules based on deep neural network," *Sustainable Energy Technologies and Assessments*, vol. 48, Art. no. 101545, 2021.

- [24] 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.
- [25] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv:2010.11929, 2020 (preprint).
- [26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
- [27] A. Salah, M. Jabrane, and I. Sebari, "A comparison of CNN, transformer, and open-vocabulary architectures for real-time photovoltaic defect detection using UAV thermal imagery," ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XI-3-2026, 2026, doi: 10.5194/isprs-annals-XI-3-2026-261-2026.
- [28] S. Gallmetzer, M. Sondoqah, E. Turri, L. Koester, A. Louwen, and D. Moser, "Redefining failure detection in PV systems: A comparative study of GPT-4o and ResNet's computer vision in aerial infrared imagery analysis," EPJ Photovolt., vol. 16, 2025, doi: 10.1051/epjpv/2025010.
- [29] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), PMLR, vol. 54, pp. 1273–1282, 2017.
- [30] X. Wang, B. Yang, Q. Liu, T. Jin, and C. Chen, "Collaboratively diagnosing IGBT open-circuit faults in photovoltaic inverters: A decentralized federated learning-based method," in Proc. IECON 2021—47th Annu. Conf. IEEE Ind. Electron. Soc., 2021, doi: 10.1109/IECON48115.2021.9589794.
- [31] F. Xu and Z. Wu, "Federated contrastive self-supervised pre-training for photovoltaic panels defect detection," in Proc. SPIE, 4th Int. Conf. Image Process. Intell. Control (IPIC 2024), vol. 13250, Art. no. 132502Z, 2024, doi: 10.1117/12.3038604.