International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

INTERPRETABLE MACHINE LEARNING FOR PHOTOVOLTAIC DEFECT SCREENING: A REPRODUCIBLE ELECTROLUMINESCENCE BENCHMARK

Khandoker Hoque *

School of Engineering, San Francisco Bay University, Fremont, CA 94539, USA
* Khandoker Hoque

ORCID Details
Khandoker Samiul Hoque: https://orcid.org/0009-0008-0756-2395

International Journal of Science and Research Archive, 2026, 20(02), 371–378

Publication history: Received on 01 July 2026; revised on 09 August 2026; accepted on 11 August 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.20.2.1601

ABSTRACT

Electroluminescence (EL) imaging can reveal cracks, inactive regions, and other abnormalities in photovoltaic (PV) cells, but manual interpretation is labor intensive and depends on specialist judgment. In this study, I evaluate transparent image features as a reproducible screening baseline. I use the public ELPV dataset, which contains 2,624 normalized grayscale cell images from 44 modules with expert-assigned defect probabilities, and resize each image to 96 x 96 pixels. I compare three representations: 62 interpretable intensity and gradient features, 324 histogram-of-oriented-gradient features, and their 386-feature combination. Using stratified five-fold cross-validation, I assess class-balanced logistic models for two prespecified outcomes: any annotated defect and high-confidence defect. For any annotated defect, the combined representation achieves mean accuracy 0.721, F1 0.670, receiver-operating-characteristic area under the curve (ROC AUC) 0.778, and precision-recall area under the curve 0.751. For high-confidence defects, the interpretable representation produces the highest mean F1, 0.638, and ROC AUC, 0.816. On the primary outcome, out-of-fold F1 is 0.712 for monocrystalline and 0.639 for polycrystalline cells. I use standardized coefficients to connect predictions with observable intensity dispersion, gradient tails, and dark-pixel fractions. I present this benchmark as an auditable basis for triage and method comparison, not as autonomous disposition, electrical qualification, or field deployment. Module-grouped external validation and correlation with calibrated electrical measurements remain necessary.

Keywords: Photovoltaic Reliability; Electroluminescence Imaging; Defect Screening; Interpretable Machine Learning; Reproducibility; Model Validation

1 INTRODUCTION

Reliable PV deployment depends on the ability to identify manufacturing defects, transport damage, installation damage, and degradation before they produce unacceptable safety, performance, or financial consequences. The U.S. Department of Energy identifies the testing of PV modules, inverters, and systems, the identification of failure mechanisms, and the development of reliability tests and standards as important components of safer and more reliable solar deployment [1]. EL imaging contributes to this objective because it exposes electrically inactive areas, microcracks,

broken interconnections, and other structures that are difficult or impossible to observe through ordinary visual inspection [2-6].

Despite its value, EL interpretation is not a trivial classification task. Defect morphology can be faint, spatially irregular, or confounded by normal material texture. Polycrystalline cells can exhibit strong background variation, while monocrystalline cells typically have a more homogeneous substrate. The operational meaning of a visible abnormality is also not binary: an image can contain a weak indication, a high-confidence defect, or a surface pattern that does not translate directly into measured power loss. These properties create three risks for automated analysis. First, a model may learn module- or material-specific texture instead of defect morphology. Second, a high aggregate accuracy can conceal poor recall for the defect class. Third, a visually plausible prediction can be mistaken for an engineering qualification decision even though no electrical or lifetime evidence was used.

Deep convolutional networks have achieved strong results on EL imagery [4-6], but they are not the only useful research target. Laboratories, manufacturers, and researchers also need transparent baselines that can be reproduced on modest hardware, inspected for subgroup behavior, and challenged through ablation. Deitsch et al. demonstrated both support-vector and convolutional approaches on cell-level EL images [2]. The ELPV benchmark subsequently enabled extensive comparison across classification and segmentation methods [3,4]. More recent work has advanced semantic segmentation, transfer learning, and high-throughput module inspection [5,6]. However, the literature still benefits from studies that report the performance of interpretable feature families, preserve annotation uncertainty, and make limitations explicit.

In this paper, the term severity classification is used narrowly for the dataset's expert-annotation confidence strata; it is not a claim that the model estimates electrical power loss, safety risk, or field-failure severity.

I designed this paper as the benchmark component of a broader, physically grounded PV-diagnostics program. Here, I compare transparent feature families, label definitions, and material subgroups. I do not use inverter-test evidence or convert EL annotations into electrical qualifications; those links require the separate acquisition-shift and physical-validation work described in the companion study.

I therefore address four questions:

1. How accurately can computationally modest, transparent image features separate functional cells from cells with any annotated defect?
2. Does performance change when the positive class is restricted to high-confidence defects?
3. Does combining interpretable features with histogram-of-oriented-gradient (HOG) descriptors improve discrimination?
4. Are results consistent across monocrystalline and polycrystalline cells?

I do not present this contribution as a claim of production readiness. It is a reproducible benchmark that connects model performance to observable image properties, reports material-type differences, and defines the validation work required before a screening model could influence engineering decisions.

2 RELATED WORK

EL imaging has become an important PV diagnostic because forward-biased cells emit near-infrared radiation whose spatial distribution reflects electrical activity. Cracks, disconnected regions, and defective interconnections may appear as dark or discontinuous structures. Buerhop-Lutz et al. introduced a benchmark for visual identification of defective cells, and Deitsch et al. reported automatic cell classification using conventional and convolutional methods [2,3]. The associated ELPV data contain 2,624 cell images from 44 modules, normalized for size and perspective and annotated with probabilities reflecting expert confidence.

Subsequent work expanded the automated workflow. Deitsch et al. addressed the segmentation of individual cells from uncalibrated module-level EL images, showing that accurate extraction is a necessary upstream step for scalable classification [4]. Akram et al. proposed a convolutional framework for automatic defect detection and reported substantial benefits from data augmentation [5]. Pratt et al. used U-Net semantic segmentation to identify and quantify cracks, inactive regions, and gridline defects, illustrating the value of pixel-level outputs for accelerated-stress and quality analysis [6]. These studies show that automated EL analysis can move beyond image-level labels toward localization and quantification.

I deliberately address a narrower question in this study. HOG descriptors summarize local gradient orientation and have a long history in visual recognition [7]. Logistic modeling provides a transparent linear decision function after

feature standardization and allows the direction and relative magnitude of coefficients to be inspected [8]. Although this combination cannot represent every nonlinear crack morphology, it offers a low-complexity benchmark whose inputs and failure modes can be audited. I also emphasize PR AUC because defect data are imbalanced and because ROC AUC alone can obscure performance on the operationally important positive class [9].

3 MATERIALS AND METHODS

3.1 Dataset and provenance

I used the ELPV repository maintained by ZAE Bayern [10]. I analyzed 2,624 normalized 8-bit grayscale images, each 300 x 300 pixels, extracted from 44 PV modules. The image license is Creative Commons Attribution-Noncommercial-Share Alike 4.0; the accompanying code is distributed under Apache 2.0.

The label file contains an image path, an expert-annotated defect probability, and a cell type. The probability distribution in the analyzed revision was 1,508 images at 0, 295 at 1/3, 106 at 2/3, and 715 at 1. The material distribution was 1,074 monocrystalline and 1,550 polycrystalline cells.

Two binary outcomes were defined before model fitting:

- Any annotated defect: probability greater than 0 (1,116 positive; 1,508 negative).
- High-confidence defect: probability at least 2/3 (821 positive; 1,803 negative).

The first task tests sensitivity to all nonzero annotations. The second reduces the influence of low-confidence cases while retaining both the 2/3 and 1 categories. Neither label is interpreted as a direct measure of electrical power loss.

3.2 Image preprocessing

I converted each image to grayscale and resized it to 96 x 96 pixels using bilinear interpolation. Pixel values were scaled to the interval [0,1]. No random augmentation, denoising, histogram equalization, or test-time transformation was used. This restrained preprocessing was selected so that model behavior could be linked to the original intensity and gradient structure.

3.3 Feature configurations

I compared three prespecified configurations.

Interpretable features (62 variables). Fourteen global variables summarized mean intensity, standard deviation, five intensity quantiles, dark- and bright-pixel fractions, gradient magnitude, and center-to-border contrast. The image was then divided into a 4 x 4 grid. Mean intensity, intensity standard deviation, and mean gradient magnitude were computed for each region, producing 48 spatial variables.

HOG features (324 variables). The 96 x 96 image was divided into 6 x 6 cells of 16 x 16 pixels. Each cell was represented by a nine-bin unsigned gradient-orientation histogram normalized by its Euclidean norm.

Combined features (386 variables). The interpretable and HOG vectors were concatenated.

3.4 Model and cross-validation

I used a class-balanced logistic model for each configuration. I standardized features using the mean and standard deviation calculated from the training portion of each fold. I optimized the model with the Adam algorithm for 240 full-batch epochs, a learning rate of 0.035, and L2 coefficient of 0.001. Positive and negative observations received inverse-frequency weights within each training fold.

I used stratified five-fold cross-validation. I fitted all preprocessing statistics and model parameters inside the training fold. I used the same deterministic seed (42) for fold assignment and model initialization. A threshold of 0.5 converted probabilities to binary predictions.

3.5 Evaluation measures

I prespecified accuracy, balanced accuracy, precision, recall, F1, ROC AUC, and PR AUC. F1 was the primary model-selection measure because it balances precision and recall for the defect class. After choosing the configuration with the highest mean cross-validated F1 for each task, I combined the out-of-fold predictions to calculate a confusion matrix and cell-type subgroup results.

I used standardized coefficients from a model fitted to all observations only for descriptive interpretation. Coefficient magnitude does not establish causality, and correlated image features can redistribute weight among themselves.

4 RESULTS

4.1 Cross-validated model comparison

For any annotated defect, the combined configuration achieved the highest mean F1 (0.670 +/- 0.018), ROC AUC (0.778 +/- 0.013), and PR AUC (0.751 +/- 0.012). The interpretable-only configuration was close, with mean F1 of 0.659 +/- 0.040 and ROC AUC of 0.768 +/- 0.033. HOG alone was weaker, with mean F1 of 0.634 +/- 0.029.

For high-confidence defects, the interpretable configuration achieved the highest mean F1 (0.638 +/- 0.027), ROC AUC (0.816 +/- 0.016), and PR AUC (0.732 +/- 0.030). The combined model produced a similar F1 of 0.634 +/- 0.028 but lower ROC AUC of 0.798 +/- 0.021. These results indicate that the value of HOG descriptors depends on the label definition; combining every feature family did not uniformly improve performance.

I summarize the cross-validated comparison in Table 1.

Table 1: Stratified five-fold cross-validation results

Task	Feature configuration	Accuracy	Balanced accuracy	Precision
Any annotated defect	Interpretable	0.712 +/- 0.035	0.704 +/- 0.035	0.663 +/- 0.043
Any annotated defect	HOG	0.694 +/- 0.018	0.685 +/- 0.021	0.645 +/- 0.021
Any annotated defect	Combined	0.721 +/- 0.011	0.714 +/- 0.013	0.675 +/- 0.013
High-confidence defect	Interpretable	0.753 +/- 0.022	0.738 +/- 0.022	0.591 +/- 0.032
High-confidence defect	HOG	0.732 +/- 0.016	0.704 +/- 0.025	0.565 +/- 0.023
High-confidence defect	Combined	0.761 +/- 0.010	0.734 +/- 0.021	0.609 +/- 0.010
Task	Feature configuration	Recall	F1	ROC AUC
Any annotated defect	Interpretable	0.656 +/- 0.039	0.659 +/- 0.040	0.768 +/- 0.033
Any annotated defect	HOG	0.625 +/- 0.046	0.634 +/- 0.029	0.743 +/- 0.011
Any annotated defect	Combined	0.666 +/- 0.032	0.670 +/- 0.018	0.778 +/- 0.013
High-confidence defect	Interpretable	0.695 +/- 0.039	0.638 +/- 0.027	0.816 +/- 0.016
High-confidence defect	HOG	0.628 +/- 0.051	0.594 +/- 0.034	0.764 +/- 0.029
High-confidence defect	Combined	0.661 +/- 0.050	0.634 +/- 0.028	0.798 +/- 0.021

Task	Feature configuration	PR AUC
Any annotated defect	Interpretable	0.748 +/- 0.031
Any annotated defect	HOG	0.716 +/- 0.008
Any annotated defect	Combined	0.751 +/- 0.012
High-confidence defect	Interpretable	0.732 +/- 0.030
High-confidence defect	HOG	0.679 +/- 0.048
High-confidence defect	Combined	0.722 +/- 0.035

4.2 Primary-task error analysis

The combined model's out-of-fold predictions for any annotated defect produced 1,150 true negatives, 358 false positives, 373 false negatives, and 743 true positives. Overall out-of-fold accuracy was 0.721 and F1 was 0.670. The near balance between false positives and false negatives shows that the threshold did not simply maximize one error type, but both error counts remain too large for unreviewed qualification decisions.

False negatives are particularly important in a reliability-screening context because a missed defect could proceed to later manufacturing, procurement, or installation stages. False positives also have operational cost because they trigger additional review or testing. The appropriate threshold should therefore be selected using the laboratory's downstream consequences and should not be assumed from the default 0.5 used in this benchmark.

4.3 Material-type subgroup performance

Performance differed by cell type. For monocrystalline cells, the primary model achieved out-of-fold F1 of 0.712, ROC AUC of 0.803, and PR AUC of 0.807. For polycrystalline cells, F1 was 0.639, ROC AUC was 0.766, and PR AUC was 0.720. Polycrystalline texture may obscure or resemble gradient and intensity patterns associated with defects, but the present experiment cannot establish the mechanism. The result demonstrates that aggregate performance should not be treated as proof of equal validity across material types.

4.4 Interpretable coefficient patterns

The largest absolute standardized coefficients in the full-data interpretable model included the 90th percentile of gradient magnitude, global intensity standard deviation, the fraction of pixels darker than 0.20, the lower intensity quartile, gradient dispersion, and median intensity. Several lower-image spatial gradient and center-region intensity variables also contributed. Collectively, the coefficients indicate that defect annotation was associated with changes in darkness, contrast, and localized edge structure. Because the variables are correlated, these coefficients should be read as descriptive model signals rather than independent physical effects.

Figure 1 provides a visual comparison of the three feature configurations.

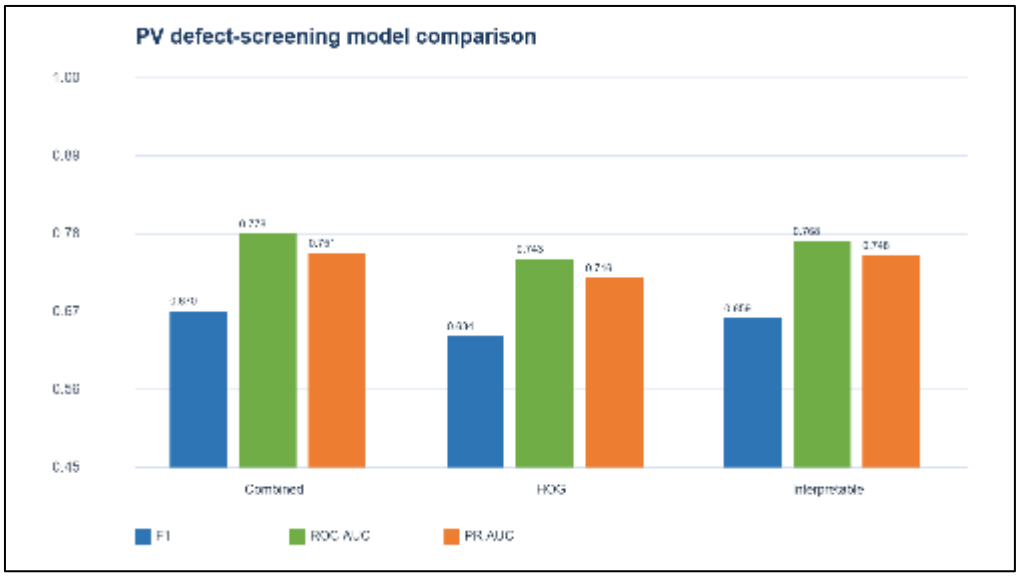

Figure 1: Cross-validated comparison of interpretable, HOG, and combined PV screening models.

Figure 2 shows the out-of-fold error counts for the primary task.

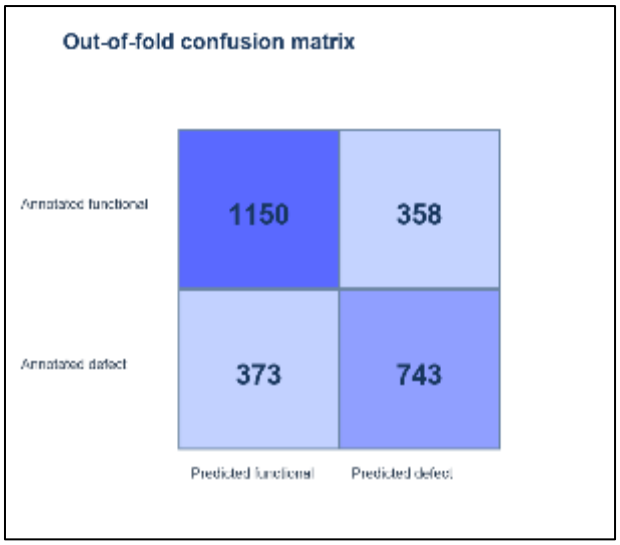

Figure 2: Out-of-fold confusion matrix for the primary any-annotated-defect task.

Figure 3 illustrates representative cells across the annotation-probability categories.

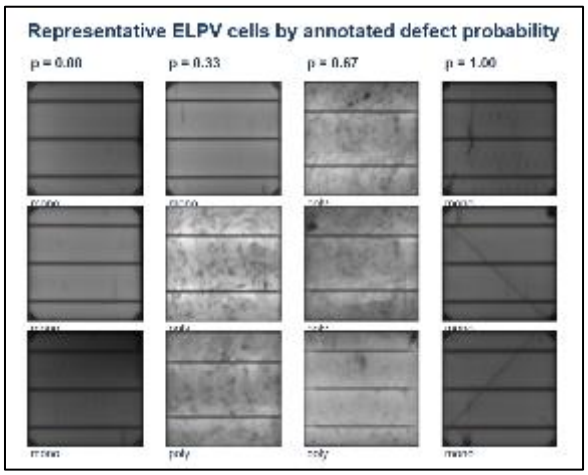

Figure 3: Representative ELPV cells across expert-assigned annotation probabilities.

5 DISCUSSION

My results provide qualified answers to the four research questions. First, transparent image features provided useful discrimination, but the achieved F1 values were moderate rather than deployment-grade. Second, redefining the target changed both the best feature configuration and the balance between precision and recall. Third, HOG added value for the broad any-defect task but not uniformly for the high-confidence task. Fourth, the material-type analysis revealed a meaningful performance gap.

These findings are useful precisely because they resist an inflated interpretation. A model that achieves ROC AUC near 0.78 on an image-level split may be suitable as a baseline, triage aid, or research comparator. It is not evidence that the model can replace EL analysts, predict power loss, or generalize to a different camera, module architecture, factory, field environment, or defect taxonomy. Deep architectures have reported higher classification accuracy in related studies [5], but comparisons across different splits, label definitions, augmentation procedures, and leakage controls are not direct.

The analysis also shows why interpretable baselines matter. The high-confidence task was best served by the 62-variable model, and its strongest coefficients corresponded to understandable intensity and gradient properties. Such a model can be audited more easily than a large end-to-end network and can reveal when a more complex approach is not buying consistent improvement. In a laboratory workflow, a transparent model could prioritize images for review, compare data distributions between batches, or support method-development experiments while final disposition remains tied to qualified procedures and electrical evidence.

At system scale, reliable PV inspection can reduce avoidable uncertainty in renewable-generation assets serving increasingly electrified and compute-intensive loads. I do not use this study to model artificial-intelligence data-center demand or grid adequacy. The relevance is narrower and technical: interpretable screening can support more consistent maintenance and validation decisions when predictions remain traceable to image evidence and are paired with calibrated electrical tests and qualified review.

Limitations

The public labels file does not expose module identifiers. Consequently, module-grouped cross-validation could not be performed, and images from the same source module may occur in both training and test folds. Shared module characteristics could therefore inflate performance. This is the most important methodological limitation.

The defect probabilities represent expert image annotations. They are not direct measurements of current-voltage degradation, insulation failure, thermal risk, lifetime, or field failure. The labels also do not identify a mutually exclusive defect mechanism. The present binary tasks therefore evaluate visual screening, not engineering severity or causal failure diagnosis.

The images were acquired and normalized within one public benchmark. No independent external dataset was used. Resizing can attenuate fine microcracks, and hand-engineered features cannot capture every morphology. Cross-validation estimates uncertainty due to fold composition but not uncertainty due to dataset shift. Finally, the subgroup comparison was descriptive and did not control for module, acquisition, or defect-distribution differences.

Future Work

The next study should collect or obtain module identifiers and perform grouped or leave-one-module-out validation. External testing should include images from different cameras, exposure settings, module designs, laboratories, and field conditions. Defect-localization labels would permit evaluation of whether the model responds to the actual abnormal region rather than background texture.

Most importantly, image predictions should be correlated with electrical and reliability outcomes, including I-V curve parameters, inactive-area estimates, insulation and leakage findings where relevant, accelerated-stress history, and longitudinal power loss. A calibrated risk model could then be evaluated as one layer in a traceable engineering workflow. Future models should compare transparent baselines with compact convolutional or transformer architectures under identical grouped splits and should report calibration, subgroup performance, threshold tradeoffs, and inference cost.

6 CONCLUSION

In this study, I establish a transparent and reproducible baseline for PV-cell defect screening in EL imagery. Combined intensity, spatial, and gradient features performed best for the broad any-defect task, while interpretable features alone performed best for high-confidence defects. Moderate F1 values, substantial false-positive and false-negative counts, and lower performance on polycrystalline cells show that the benchmark is not a deployment claim. Its value is an auditable foundation for module-grouped validation, external testing, defect localization, and eventual linkage between image evidence and electrical reliability outcomes.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

I thank the maintainers of the public datasets and open research resources used in this study. I received no external funding, and I did not use employer-confidential data, client data, or proprietary laboratory records.

Disclosure of conflict of interest

I, Khandoker Samiul Hoque, declare no competing financial or non-financial interests relevant to this work. My employment did not sponsor the study.

Author Contributions

As the sole author, I was responsible for conceptualization, methodology, software, validation, formal analysis, visualization, data curation, and writing - original draft and review.

Data Availability

I identify the public source datasets, deterministic analysis scripts, derived metrics, and figure files in the accompanying reproducibility materials. I record repository revisions, seeds, preprocessing rules, and experimental parameters in the Methods and supplementary files. Reuse remains subject to the source repositories' licenses and terms.

REFERENCES

[1] U.S. Department of Energy. Reliability and Safety. Solar Energy Technologies Office. https://www.energy.gov/cmei/systems/reliability-and-safety.

[2] Deitsch S, Christlein V, Berger S, Buerhop-Lutz C, Maier A, Gallwitz F, Riess C. Automatic classification of defective photovoltaic module cells in electroluminescence images. Solar Energy. 2019;185:455-468. doi:10.1016/j.solener.2019.02.067.

[3] Buerhop-Lutz C, Deitsch S, Maier A, Gallwitz F, Berger S, Doll B, Hauch J, Camus C, Brabec CJ. A benchmark for visual identification of defective solar cells in electroluminescence imagery. 35th European Photovoltaic Solar Energy Conference and Exhibition. 2018:1287-1289. doi:10.4229/35thEUPVSEC20182018-5CV.3.15.

[4] Deitsch S, Buerhop-Lutz C, Sovetkin E, Steland A, Maier A, Gallwitz F, Riess C. Segmentation of photovoltaic module cells in uncalibrated electroluminescence images. Machine Vision and Applications. 2021;32:84. doi:10.1007/s00138-021-01191-9.

[5] Akram MW, Li G, Jin Y, Chen X, Zhu C, Ahmad A. CNN based automatic detection of photovoltaic cell defects in electroluminescence images. Energy. 2019;189:116319. doi:10.1016/j.energy.2019.116319.

[6] Pratt L, Govender D, Klein R. Defect detection and quantification in electroluminescence images of solar PV modules using U-Net semantic segmentation. Renewable Energy. 2021;178:1211-1222. doi:10.1016/j.renene.2021.06.086.

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005:886-893. doi:10.1109/CVPR.2005.177.

[8] Cox DR. The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B. 1958;20(2):215-242.

[9] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.

[10] ZAE Bayern. ELPV Dataset: A Benchmark for Visual Identification of Defective Solar Cells in Electroluminescence Imagery. GitHub repository. https://github.com/zae-bayern/elpv-dataset.