Downloaded from orbit.dtu.dk on: May 04, 2024

DTU Library

Improving Deep Learning-Based Defect Classification in Solar Cells using Conformal Prediction

Thomsen, Vitus B.; Mantel, Claire; Benatto, Gisele Alves dos Reis; Engsig-Karup, Allan Peter; Forchhammer, Søren

Published in:
Proceedings of 50th IEEE Photovoltaic Specialists Conference

Link to article, DOI:
10.1109/PVSC48320.2023.10360000

Publication date:
2023

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Thomsen, V. B., Mantel, C., Benatto, G. A. D. R., Engsig-Karup, A. P., & Forchhammer, S. (2023). Improving Deep Learning-Based Defect Classification in Solar Cells using Conformal Prediction. In *Proceedings of 50th IEEE Photovoltaic Specialists Conference* IEEE. https://doi.org/10.1109/PVSC48320.2023.10360000

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Improving Deep Learning-Based Defect Classification in Solar Cells using Conformal Prediction

Vitus B. Thomsen, Claire Mantel, Gisele A. dos Reis Benatto, Allan P. Engsig-Karup, Søren Forchhammer

Technical University of Denmark, Kongens Lyngby, 2800, Denmark

***Abstract* — Deep learning-based approaches have become popular for automatically detecting defects in electroluminescence images of solar cells. However, deep learning methods are those that require the most training data among machine learning approaches. Thus, the data available to train such models is currently a bottleneck for their performances due to expensive and possibly inaccurate labeling. To address this problem, we propose to use a model comprising a standard deep learning classifier to which we add conformal prediction. The model calculates a degree of confidence on new predictions and can send low-confidence predictions for human expert labeling in an uncertainty-aware active learning loop. In tests with a limited-size data set, using the conformal model to select and classify high-confidence samples yields significantly higher performance compared to the standard deep learning classifier, as the F1 score increases from 0.44 to 0.62 while only leaving out 9.4% of predictions as low-confidence that need human assessment for validation and model update, demonstrating the effectiveness of the framework.**

I. Introduction

Inspection of photovoltaic (PV) modules is crucial to ensure optimal power output. Electroluminescence (EL) imaging of PV cells is an effective way of performing inspection, as it allows for greater detail than when using thermal or visible imaging only, acting like an "X-ray" image of the PV device [1]. In recent years, there has been extensive research in using machine learning, especially convolutional neural networks (CNN), on EL images for detecting defective cells [2]-[4]. A major limitation of the performance of learning-based models is currently the data available. Indeed, training a model requires large datasets whose quality depends on the variety and distribution of defects. For supervised learning (the most common), models need to be trained on labeled datasets. These labels are annotations from human experts which make them both very time consuming to create and prone to containing potential errors (inaccuracies, inconsistencies, or omissions).

In machine learning, the idea of conformal prediction allows for uncertainty quantification of a model. Conformal prediction is a general framework for constructing prediction intervals for machine learning models with a guaranteed level of accuracy, regardless of the distribution of the underlying data, that can be implemented in a general way and comes with a wide array of applications [5][6].

Another important framework in machine learning is the concept of active learning, which can be used to define an algorithm driven training loop that relies on a machine learning model and some measure of information content that can be used to ask a human expert for input on new instances that are of most value to further improve the machine learning model. During active learning, the machine learning model is then conveniently retrained to be updated using newly labeled data, allowing for optimal model improvement over time while keeping the cost of labeling data to a minimum as only a minimal amount of new images is seen by an expert for labeling [7][8]. Active learning can be done in several ways, and in this work conformal prediction is used to define an uncertainty-aware active learning loop for determining the low-confidence data to be labeled by a human. This idea is also explored in other works, e.g. [9].

In this work, we develop and test an uncertainty-aware active learning framework for PV classification based on conformal prediction and apply it to a deep learning model. The objective is that this framework will make it easier to improve the accuracy of PV classification models as more data becomes

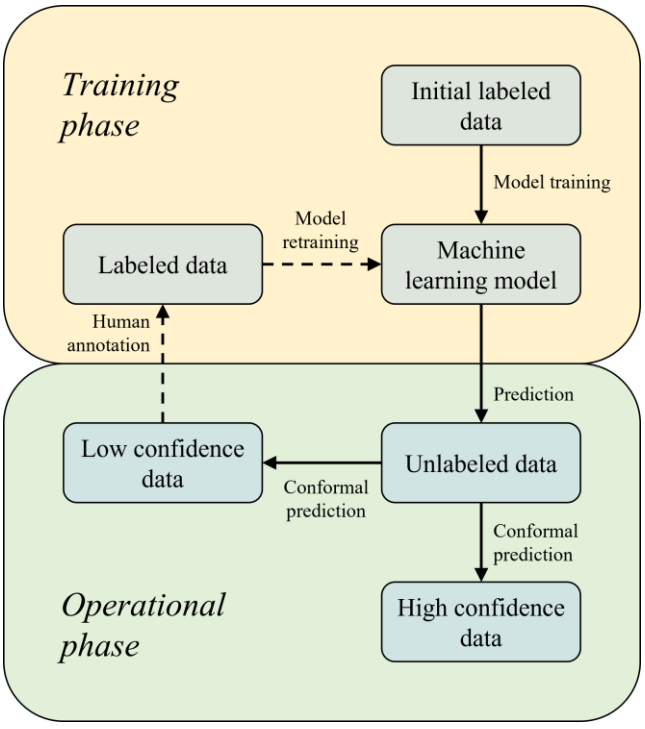

Fig. 1. Illustration of the proposed uncertainty-aware active learning framework based on conformal prediction. The dashed arrows indicate steps that are not directly considered in this work.

available, hence dealing with the challenges of doing manual annotations and improving model accuracy at the same time.

Fig. 1 illustrates the proposed active learning loop. Initially, a deep learning-based classification model is trained using existing labeled data (i.e., EL images of solar cells). Whenever new data is collected, the model is used to predict if each cell is defective or not. In this work, we also predict the type of defect, although the framework can be applied to simple binary classification as well. Using conformal prediction, each prediction is then identified as either high or low confidence. The high-confidence predictions are trusted, while the remaining samples are handed off to a human annotator (possibly along with the corresponding low-confidence predictions) who will manually go through these samples and correctly label the data. The newly labeled data is then used to retrain the model. The benefits of the active learning framework are thus twofold. The first benefit is that the predictions can be made more accurate and hence more trustworthy when restricted to the high confidence samples. The second benefit is that we can choose the examples for manual labeling that are most useful for retraining the model, and thus the model can be improved over time in an efficient manner. This work focuses on setting up the conformal prediction framework and the first of these benefits, i.e., optimizing the model accuracy on the high-confidence samples. We leave the annotation and retraining steps to be explored in more detail in future work.

We propose a simple way of splitting the predictions into high and low confidence, as well as a method of tuning the framework such that the high confidence predictions are accurate while having as few low confidence predictions as possible.

II. Methodology

A. Data filtering and preprocessing

The dataset consists of EL images collected from different solar farms using a high-resolution silicon-based detector camera at nighttime [10]. A perspective correction algorithm has been applied to the panel images, followed by local brightness normalization and algorithmic separation into individual cells [11]. Four different types of defects are considered: cracks of modes A, B, and C, as well as finger failures, as defined in [1]. Examples of defective cells are seen in Fig. 2. Each defective cell has been annotated with the type of defect and its location within the image as a binary mask.

Some cells in the dataset present potential induced degradation (PID), but this defect was not among the considered defects and those cells were excluded from the dataset by thresholding the mean pixel intensity. Furthermore, the cell images were checked for correct content, and those incorrectly cropped were discarded as well.

A small number of cells are annotated with more than one defect type. Since this work is focused on simple multiclass classification, where each cell belongs to only one of several classes (as opposed to multilabel classification, where a cell may belong to multiple classes at the same time), a single defect type is assigned to each of the multi-labeled cells. The assigned class is chosen using the following prioritized list based on how important or severe the four defect types are deemed, in order from highest to lowest priority: Crack C, crack B, finger failure, and crack A.

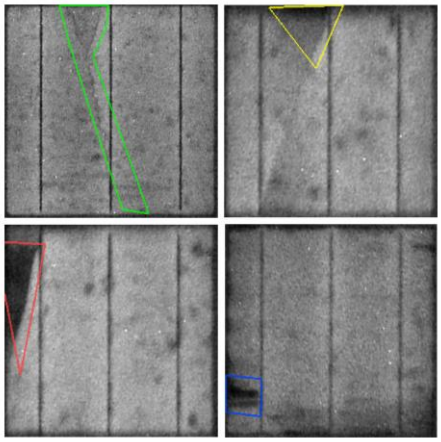

Fig. 2. Examples of cells with annotations from different defective classes. The defect types are crack A (top left), crack B (top right), crack C (bottom left), and finger failure (bottom right).

As the labeling was done on full modules, the marked locations of the defects were also split into cells. Special attention was made to avoid false positives from manual marking of a defect in one cell jutting out on neighboring cells.

After the data filtering, the dataset contains 35969 images of cells split into five classes – one class for the non-defective cells, and one class for each of the four defect types. The distribution of the five classes is seen in Table I. It is seen that the dataset is highly imbalanced, with only 993 cells in total labeled as defective.

TABLE I. Distribution of Classes

Class	Number of samples
No defect	34976
Crack A	130
Crack B	232
Crack C	124
Finger failure	507
Total	35969

B. Model selection

The dataset is split into a training, validation, and testing set in the proportions 60:20:20 with class stratification. The basic model is a classifier using a CNN, which is trained on the training set. Model selection is done using the macro-averaged F_1 score as a performance metric, evaluated on the validation set. The usual F_1 score for a single class is defined as

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (1)$$

where precision and recall are given by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (3)$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The macro-averaged F_1 score is then given by the mean of the F_1 scores on the four defect classes (the no defect class is excluded).

After initial tests with different network architectures, a transfer learning model based on the VGG-13 architecture [12] with batch normalization is chosen. The first and last layers have been slightly modified, such that the network accepts grayscale images as input rather than RGB images, and such that the network has five output neurons, corresponding to the five classes. All layers of the network are trained simultaneously.

Model selection is done in several stages. A few different regularization techniques are tried to avoid overfitting, one at a time. Label smoothing is a technique used to avoid the model becoming overly confident in its predictions [13]. It is found that adding a small amount of label smoothing (smoothing level = 0.05) improves the F_1 score. Weight decay (L2 regularization) was also tried but was not found to improve the model. Hyperparameters (the learning rate and its decay rate) are also tuned by a small grid search. The networks are trained using the cross-entropy loss function. The Adam optimizer [14] is used with a learning rate of 0.001, which is set to exponentially decay by a factor of 0.995 after every epoch. Training is done for 300 epochs using a batch size of 64. Weighted random sampling is used during training, so the model sees roughly the same number of images from each class.

Due to the small number of images in the defective classes, data augmentation is used to introduce more variation to the training data. This is done by randomly applying transformations to each image on-the-fly during training, each with probability 1/2. The transformations that are found to improve the model are: horizontal/vertical flipping, 90-degree rotations, small rotations (±2°), brightness and contrast adjustments (±25%), gamma adjustments (γ chosen log-uniformly from [2/3, 3/2]), and random cropping, which is done in such a way that the bounding box of the defective region is kept (mostly) within the cropped image.

C. Conformal prediction

Conformal prediction is a general framework for constructing prediction intervals for machine learning models [5][6]. For a classifier, this allows for outputting a prediction set of classes rather than a single predicted class. A user-defined α allows for obtaining a set $C_{1-\alpha}$ that will contain the true class with probability at least 1 - α. Due to our limited amount of data, we here apply the cross-conformal prediction algorithm known as cross-validation+ [15]. In summary, the algorithm works by splitting the training set into K folds (we use $K = 5$), then training K different classifier networks, each using only $K - 1$ of the folds as a proper training set. For each of the samples in the remaining fold, a certain conformity score function is computed by comparing the raw model outputs to the true labels. These scores are referred to as the hold-out scores. When making predictions on new data, the prediction set is constructed as the set of classes such that the resulting conformity score of the new sample is smaller than $(1 - \alpha)(n + 1)$ corresponding hold-out scores, aggregated over all K classifiers. This method ensures that we meet the above-mentioned coverage guarantee of 1 - α. For a more detailed description of the algorithm, we refer to the original paper [15].

In the proposed active learning framework, the predictions are split into high and low confidence. The idea is that predictions for which the model has low confidence will be sent to a human expert for annotation and/or verification. In this study, we consider a prediction as high-confidence if its prediction set is a singleton, i.e., it contains exactly one class, and low-confidence otherwise (0 or ≥2 classes). Note that the prediction set can contain both the no defect class and any of the four defect classes.

We aim at maximizing jointly these two aspects: (i) The performance of the model when restricted to the high-confidence predictions, and (ii) The proportion of predictions that are considered high-confidence. With both these goals in mind, we define the following metric for evaluating the performance of the conformal model:

$$\tilde{F}_1 := \frac{N^{\text{conf}}}{N^{\text{val}}} F_1^{\text{conf}} \quad (4)$$

where F_1^{conf} is the macro averaged F_1 score evaluated only on the trusted high-confidence predictions, N^{conf} is the number of high-confidence predictions, and N^{val} is the size of the validation set, i.e., the total number of predictions we make. We tune the conformal model by optimizing $\tilde{F}_1$ over the following two parameters: the significance level α (i.e., to what level of certainty should the prediction sets be constructed) and the number of epochs of training. The optimal set of values for these two parameters is determined by a simple grid search.

On a somewhat technical note, the cross-validation+ algorithm inherently involves randomness to construct the conformity scores and thus the prediction sets. To get more robust results, we generate 50 realizations of the prediction sets for each combination of parameters by using 50 different seeds for the random number generator, then average the relevant scores over these.

Both the conformal model and a 'standard' classifier are then retrained using the union of the training set and the validation set as the new training set, then evaluated on the test set.

III. Results

As a baseline model, a standard classifier is trained on the combined training/validation set and evaluated on the test set. The model is trained for 300 epochs. The macro-averaged F_1 score of the model during training is seen in Fig. 3. The final F_1 score obtained is 0.44, although the exact value fluctuates a lot between epochs. A notable observation is that after training for approximately 150 epochs, the F_1 score is not found to increase any further. This is explained by the fact that the precision is increasing with the epoch number, while the recall is decreasing, and these two effects seem to cancel each other out in the F_1 score. This suggests that the model tends to become increasingly biased towards classifying cells as non-defective the longer it is trained.

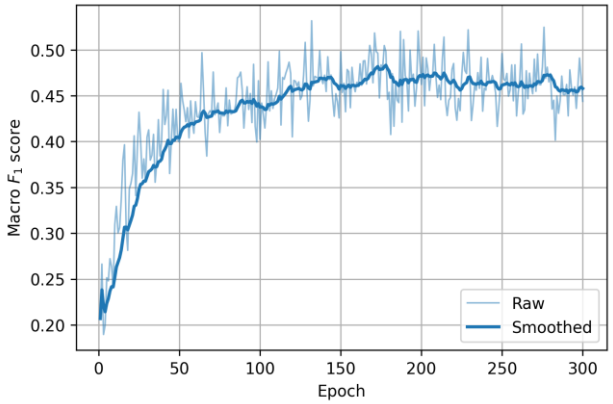

Fig. 3. Macro-averaged F_1 score of the final standard classifier, evaluated on the test set.

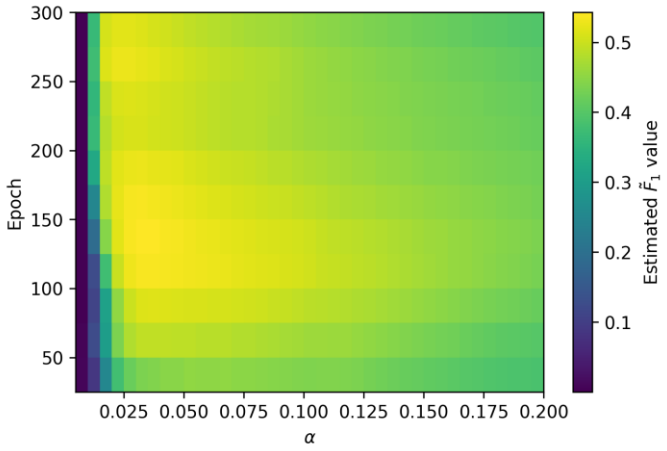

Fig. 4. Estimated $\tilde{F}_1$ values of the conformal model during the tuning process, evaluated at different α values and epoch numbers. A small amount of Gaussian smoothing has been applied in the epoch direction.

When tuning the conformal model on the validation set, we find that the optimal $\tilde{F}_1$ value occurs at α = 0.03 and after 125 epochs, with an estimated $\tilde{F}_1$ value of 0.54 (after some smoothing). The full grid search is seen in Fig. 4. In general, it is observed that lower values of α tend to give a greater $\tilde{F}_1$ value, although very small values such as α = 0.005 gives very poor results. This is mainly because very small values of α result in prediction sets that are almost always of size ≥2 to achieve 1 - α coverage. On the other hand, large values of α make it more likely for the model to give empty prediction sets.

Using the found optimal values, the conformal model is retrained on the combined training/validation set and applied to construct prediction sets for the test set. Once again, in order to get more robust results, we generate 50 realizations of the prediction sets and find the F_1^{conf} scores for each realization, however this time, we find the median of these F_1^{conf} scores and only consider the realization giving this median score. This is then considered representative of the typical outcome of the conformal model.

The confusion matrix for the standard classifier is shown in Table II, while the confusion matrix for the high-confidence predictions by the conformal model is shown in Table III (classes are abbreviated as ND = no defect, A/B/C = crack A/B/C, FF = finger failure). The key observation is that for the conformal model, the predictions are more concentrated on the diagonal of the confusion matrix. Table IV shows some key numbers summarizing the confusion matrices. We see

TABLE II. Standard Classifier

		Predicted					
		ND	A	B	C	FF	Total
Actual	ND	6907	6	17	4	61	6995
	A	18	5	2	0	1	26
	B	14	4	24	2	2	46
	C	4	1	5	13	2	25
	FF	46	1	1	2	52	102
	Total	6989	17	49	21	118	7194

TABLE III. Conformal Model

		Predicted					
		ND	A	B	C	FF	Total
Actual	ND	6416	1	7	0	20	6444
	A	8	4	0	0	0	12
	B	3	1	10	0	0	14
	C	2	0	2	9	0	13
	FF	9	0	0	0	26	35
	Total	6438	6	19	9	46	6518

TABLE IV. Metrics for the Two Classifiers

	Standard	**Conformal**
Macro F1	0.4439	0.6277
Macro precision	0.4609	0.6895
Macro recall	0.4360	0.6207
Accuracy	0.9732	0.9919
Number of predictions	7194/7194	6518/7194

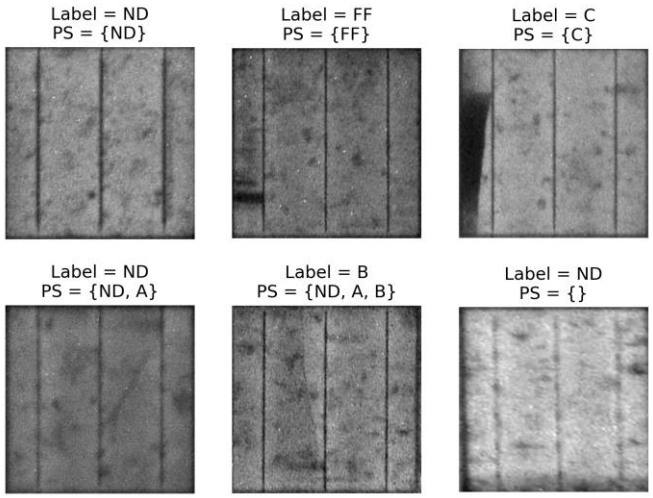

Fig. 5. Examples of test data with corresponding true labels and prediction sets (PS). The top three predictions are regarded as high-confidence since the prediction sets contain exactly one class, while the bottom three are regarded as low-confidence.

substantial improvements in F_1 score, precision, recall and accuracy for the conformal model. The price we pay for this is that 676 out of the 7194 test images (9.4%) are regarded as low-confidence, where a considerable proportion of these (18.5%) are from the defect classes.

Fig. 5 shows some examples of test data with their corresponding true labels and prediction sets, where the first three examples are high-confidence and the last three examples are low-confidence. In general, it is observed that the low-confidence samples indeed tend to be images where the true class is somewhat ambiguous, for example due to a blurry image or a crack where the type is not clear.

We remark that the conformal model has an empirical coverage of 97.9% across all prediction sets, exceeding the expected 97% (as $\alpha = 0.03$). We also remark that it is possible to use the conformal model as a 'regular' model, giving only a single prediction, by simply taking the predicted class to be the first class that would be included in the prediction set (thus using it as a sort of ensemble model over the five underlying classifiers). This method allows us to evaluate the conformal model on all test samples, not just the high-confidence ones. Doing this gives an F_1 score of 0.50, which is an improvement on the standard model but not on the conformal model restricted to high-confidence predictions. This further supports the finding that the separation into high- and low-confidence predictions is meaningful.

IV. Discussions and Future Work

The main factor limiting the performance of the models in this work is the 'real-life' quality of the dataset. Firstly, the dataset is heavily imbalanced with a lack of samples in the defective classes, which makes it difficult for a model to learn a general pattern, but also difficult to accurately assess the model performance due to a lack of test data. Secondly, the labeling has inconsistencies due to different people having labeled different parts of the dataset. Moreover, even though it was attempted to remove the factitious annotations, there may still have been some cells left with incorrect labels in the final dataset. It is certain that more data as well as relabeling the existing data can lead to further improvements of the models.

However, it was seen that the conformal prediction framework led to significant improvements when restricting to high-confidence predictions. This demonstrates that even with a dataset of non-ideal quality, the conformal prediction framework works as intended, in the sense that the high-confidence predictions by themselves are more accurate than when considering all predictions at once. The main benefit of the uncertainty-aware active learning framework is that it allows detecting the difficult samples for which the prediction can only be made with low confidence. Once these low-confidence predictions go through human evaluation to be labeled with certainty, the model is retrained to give better performance.

A significant downside of the cross-validation+ approach used in this work is its computational cost, as it requires training $K = 5$ distinct networks. This method was deemed necessary in this work due to limited data. There is also the limitation that the method cannot be easily applied to existing classifier systems without retraining. For future work, it would be relevant to explore cheaper methods such as simple split-conformal calibration [6][15] that are both computationally simpler and more readily applicable to existing systems.

Additionally, further work is required on streamlining the proposed active learning framework. The tuning process used in this work was necessary to give the desired performance improvements, and it may be revised. In a practical setting, one might want to choose the significance parameter α manually to gain control of the balance between the guaranteed level of confidence and the number of samples that must be manually labeled, depending on the application. Alternatively, one might want a fixed number of samples to be manually labeled in each iteration of the active learning loop. It is also important to stress that the parameters found in this work are by no means necessarily optimal in a general setting, as they may depend highly on the available data and the underlying deep learning model that is used.

Furthermore, while the conformal prediction framework is theoretically well-founded, there is no theoretical guarantee that the active learning framework used in this work chooses the optimal samples for learning. As such, the work done here should be seen as preliminary and a "proof-of-concept". In future work, more sophisticated ways of choosing samples for manual labeling may be considered, as has been done in other works [9]. Additionally, focus should be on incorporating tests that promote higher quality in the data. For example, it could be considered to modify the active learning loop such that we also

predict on a small amount of data that was previously deemed high-confidence to ensure that the updated model is still confident in these samples, which can work as a quality assurance step.

REFERENCES

[1] M. Köntges, S. Kurtz, C.E. Packard, U. Jahn, K.A. Berger, K. Kato, T. Friesen, H. Liu, M. Van Iseghem, J. Wohlgemuth and D. Miller, "Review of failures of photovoltaic modules," IEA-PVPS T13-01:2014, 2014.

[2] A. Bartler, L. Mauch, B. Yang, M. Reuter and L. Stoicescu, "Automated detection of solar cell defects with deep learning," *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018, pp. 2035-2039.

[3] S. Deitsch, V. Christlein, S. Berger, C. Buerhop-Lutz, A. Maier, F. Gallwitz and C. Riess, "Automatic classification of defective photovoltaic module cells in electroluminescence images," *Solar Energy*, 185, pp. 455-468, 2019.

[4] W. Tang, Q. Yang, K. Xiong and W. Yan, "Deep learning based automatic defect identification of photovoltaic module using electroluminescence images," *Solar Energy*, 201, pp. 453-460, 2020.

[5] V. Balasubramanian, S.S. Ho and V. Vovk, *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[6] A.N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification", arXiv preprint, arXiv:2107.07511, 2021.

[7] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[8] H.O. Ilhan and M.F. Amasyali, "Active learning as a way of increasing accuracy," *International Journal of Computer Theory and Engineering*, 6(6), p. 460, 2014.

[9] S. Matiz and K.E. Barner, "Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification," *Pattern Recognition*, 90, pp. 172-182, 2019.

[10] C. Mantel., F. Villebro, G.A. dos Reis Benatto, H.R. Parikh, S. Wendlandt, K. Hossain, P. Poulsen, S. Spataru, D. Sera and S. Forchhammer, "Machine learning prediction of defect types for electroluminescence images of photovoltaic panels", *Proc. SPIE*, vol. 11139, 2019, Art. no. 1113904.

[11] C. Mantel, F. Villebro, H.R. Parikh, S. Spataru, G.A. dos Reis Benatto, D. Sera, P.B. Poulsen and S. Forchhammer, "Method for estimation and correction of perspective distortion of electroluminescence images of photovoltaic panels," *IEEE Journal of Photovoltaics*, 10(6), pp. 1797-1802, 2020.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, arXiv:1409.1556, 2014.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2818-2826.

[14] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint, arXiv:1412.6980, 2014.

[15] Y. Romano, M. Sesia and E. Candes, "Classification with valid and adaptive coverage," *Advances in Neural Information Processing Systems*, 33, pp. 3581-3591, 2020.