Case Study: Utilizing Deep Learning Models for Fault Detection and Diagnosis of Photovoltaic Modules to Improve Solar Energy Constructions' O&M Activities Quality

Khuong Nguyen-Vinh[1-2], Huynh Quang Nguyen Vo[3], Khoa Nguyen-Minh[4], and Minh Hoang[5]

[1] RMIT, Saigon South, 702 Nguyen Van Linh Street, Ho Chi Minh city, Vietnam
[2] VSB – Technical University of Ostrava, 17. Listopadu 2172/15, Ostrava, Czech Republic
[3] Intel Products Vietnam, Lot I2, Road D1, Saigon Hi-Tech Park, Ho Chi Minh City, Vietnam
[4] Aalto University, Otakaari 24, Espoo, Finland
[5] University of Washington, 1410 NE Campus Parkway, Seattle, WA, USA

khuong.nguyenvinh@rmit.edu.vn, huynh.quang.nguyen.vo@intel.com

Abstract. Renewable energy sources have long been considered to be the sole alternatives to fossil fuels. Consequently, the usage of photovoltaic (PV) systems has experienced exponential growth. This growth, however, places gargantuan pressure on the solar energy industry's manufacturing sector and subsequently begets issues associated with the quality of PV systems, especially the PV module, which is the systems' most crucial component. Currently, fault detection and diagnosis (FDD) are challenging due to many factors including but not limited to requirements of sophisticated measurement instruments and experts. Recent advances in deep learning (DL) have proven its feasibility in image classification and object detection. Thus, DL can be extended to visual fault detection using data generated by electroluminescence (EL) imaging instruments. Here, the authors propose an in-depth approach to exploratory data analysis of EL data and several techniques based on supervised and unsupervised learning to detect and diagnose visual faults and defects presented in a module.

Keywords: Computer Vision, Supervised Learning, Unsupervised Learning, Deep Learning, Neural Networks.

1 Introduction

The operation and maintenance (O&M) of solar farms is an essential duty that cannot be overlooked to maintain asset functionality. Concerning the return on investment, investors are highly critical. Failures or unexpected deterioration issues will result in lower energy production and a possible loss of component warranty due to manufacturer turnover if they go undiscovered. To maximize solar energy production, all photovoltaic (PV) panels and other associated components must function properly. The ability to produce electricity and maintain the efficiency of a single panel is seriously hampered even as one of its cells is damaged [1]. The O&M services must ensure that solar energy is produced to protect investors' investments and interests.

Certain renewable energy harnessing product designs are exceedingly intricate and they require machine-like accuracy that human personnel are unable to provide. Modern technologies including Internet-of-Things (IoT), artificial intelligence (AI), robotics, and drones can lower the monitoring cost and promote efficient O&M services [2]. Data is gathered using state-of-the-art AI tools. The installation of IoT device on solar panels also facilitate the collection of vital information. This information is examined and employed to automate the PV panel maintenance process. Automation with the help of robots and AI offers a quicker and more effective means to do various tasks to enable resource saving. By employing robotics, AI, and IoT more frequently, renewable energy companies can initiate a complete swing away from unsafe energy sources (e.g., coal and oil) in favor of solar energy [3].

2 How artificial intelligence (AI) was integrated into O&M activities of solar energy systems

2.1 Background

Many solar power facilities do not have effective O&M systems. As a result, they frequently lose power that has already been generated. Modern technologies like IoT, AI with data learning, robotics, and drones can lower the monitoring cost and facilitate effective O&M services [2].

2.2 Literature Review

To stay abreast of market demands, O&M contractors are increasingly dependent on innovations and data-driven solutions. The levelized cost of electricity (LCOE) is expected to decrease as a result of innovations in O&M services [4].

The most popular deep learning algorithm that is successful for image-based diagnosis by far is the convolutional neural network (CNN). A study carried out by Shin et al. [5] utilized the same deep-learning algorithm to process images and detect faults. The crucial elements needed for defect diagnostics are automatically extracted from an image by this method. With enough training data, multiple failure modes can be learned, and the AI model can assess whether a problem exists on an undiscovered image.

Predictive maintenance is generally less effective due to high customization costs, the requirement for collecting a large number of physical variables, and the lack of a reliable Internet connection on solar farms. Additionally, the lack of a predictive component in the maintenance strategy makes it more difficult to reduce downtime costs. Statistical approaches based on data mining are now emerging as a promising solution both for failure prediction and early detection in PV panels to keep implementation costs and model complexity down [6]. While manual diagnostic of PV panels is the least expensive option, its detection accuracy is the lowest. Predictions with the help of data and machine learning are reasonably priced and offer good detection accuracy [7].

A study carried out by Huuhtanen et al. [8] indicates that when utilising CNN methods there is a deviation in fault identification in PV panels. Hence, they propose numerical tests to show the exact power curve of a working panel. Their proposed method performs better than current methods relying on basic interpolation filters. Additionally, they develop an algorithm that uses power measurement history data (time series of power measurements) of a target panel and its surrounding panels to identify defective PV panels.

Another study demonstrates the potential applications of AI to forecast PV energy generation. The proposed method is intended to serve as a module for the energy management and production scheduling of a photovoltaic power plant. The goal of the study was to create a solution that would help to deliver power efficiently based on historical and present-day data on solar radiation in real time. Different feedforward neural network designs and data input scenarios were examined, and the results provided a solution utilizing artificial intelligence to manage photovoltaic energy production [9].

Compared to any conventional method, AI algorithms have the potential to provide better, quicker, and more useful predictions of PV panel failure. Artificial neural networks, fuzzy logic, adaptive network-based fuzzy inference systems, and data mining are some of the fields of AI [10].

In the planning and improvement of renewable energy systems, AI plays an increasingly important role. Numerous AI techniques and technologies are currently being widely used in the energy industry in areas including generation forecasting, energy efficiency monitoring, energy storage, and overall energy system design [11]. Currently, complex algorithms solving differential equations are used in analytical computer codes for modeling, which is the prediction of performance and control of renewable energy systems. To create precise forecasts, this method needs a lot of processing power and time. Instead of requiring intricate rules and mathematical procedures, AI systems can learn important informational patterns. Better, faster, and more useful predictions could be made using AI than with any current method [10].

To conclude, O&M expenses are key to the economic success of the renewable energy sector and are a major factor in crucial metrics like LCOE. It is essential to find problems early on to reduce O&M costs. Since catastrophic equipment damage can be avoided and time-consuming repair schedules can be completed in advance, it is crucial to anticipate a breakdown. AI, IoT, and robotics are examples of novel technologies that can be used to do proactive PV panel maintenance, which will improve energy production efficiency and minimize energy costs.

3 Utilizing Deep Learning Models for Fault Detection and Diagnosis

3.1 Introduction and Motivation

Recent years have seen a rapid increase in the use of PV electricity. Since 2010, the cumulative annual growth rate of the PV market and the global expansion of PV capacity have increased consistently at average rates of 20% and 12% year over year, respectively [12]. Most importantly, the increasing trend of PV capacity continues despite the COVID-19 pandemic in 2020 and 2021 [13]. The exponential expansion of PV systems suggests that the world is shifting toward renewable energy sources. However, this quick expansion places significant pressure on the solar energy industry’s manufacturing sector to satisfy such high demand, resulting in several challenges related to system quality, particularly its most critical component: the PV module.

The current method for assessing the quality of PV modules mostly comprises failure detection and diagnosis (FDD) during the production phase, which is difficult due to a variety of issues. The most important aspect is that these flaws cannot be seen with the unaided eye. As a result, sophisticated apparatus and trained staff are necessary to make this procedure possible. Even with the availability of competent staff, errors made during the procedure can be significant. Due to the aforementioned high demand affecting the industrial sector, experienced employees are usually committed to long working hours. Thus, they are prone to blunders because of exhaustion and weariness.

FDD processes have long been critical in a wide range of sectors, from aircraft [14] to automobiles [15], medical equipment [16], and semiconductor devices [17]. The fundamental goal of the FDD process is to discover and diagnose errors as well as their related root causes early enough to allow fixes before further harm to the system or loss of service happens [18]. Since PV systems serve as power generators, failures in any component can adversely damage efficiency, energy production, security, and dependability if not detected and corrected quickly [19]. As a result, problems must be detected correctly during the production phase to maintain optimal efficiency and energy yield while minimizing the costs of maintenance and corrective operations.

FDD systems are classified into two types: model-based and data-driven techniques. Model-based systems incorporate domain knowledge into the system to develop a model that compares measured values of critical system parameters to reference values - commonly known as golden samples - to derive a forecast [20]. Data-driven models, on the other hand, have been constructed based on observations of input and output data [21, 22]. Thanks to the availability of massive data, substantial processing power, and the breakthrough of deep learning, the employment of data-driven systems has become increasingly appealing [23].

Recent breakthroughs in machine learning approach for image classification and pattern recognition have shown that they are completely practical for visual fault detection tasks. Few studies, however, have shown conclusive success in integrating deep learning models into the FDD process for manufacturing silicon PV modules using small datasets. As a result, the goal of this study is to propose a strategy to apply deep learning for the FDD of visual flaws on silicon PV modules intending to simplify and automate current reliability testing methods throughout the manufacturing process. The study consists of the following tasks to accomplish the objective mentioned above:

- Provide a comprehensive exploratory data analysis using a given dataset.
- Experiment with several techniques based on both supervised and unsupervised learnings to detect and diagnose visual faults and defects presented in a module using the aforementioned dataset.
- Experiment with several nuances relating to practical realizations of deep learning models.

3.2 Related Works

Exploratory Data Analysis

Exploratory data analysis is a crucial procedure that entails performing early investigations on data to find patterns, identify anomalies, test hypotheses, and validate assumptions with the aid of summary statistics and graphical representations [24]. Despite its significant role, there is seemingly a scarcity of research on how exploratory data analysis is applied in image classification as a pre-screening step. The most recent study that involves this particular subject is [25], where a comprehensive exploratory data analysis approach consisting of two primary processes for image classification is introduced. First, statistical features (e.g., mean, median, standard deviation, etc. of pixels) of the images and the regions of interest are calculated, respectively. Then the “textual” features associated with the regions of interest are extracted and described in three principal forms: statistical (calculated by the Gray Level Co-occurrence Matrix), structural, and spectral.

Supervised Learning

Supervised learning, a classical field of machine learning, is a task of learning a function that maps an input to an output based on example input-output pairs [26]. One of the primary tasks of supervised learning is classification - the same target of the study.

Considering the fault detection and diagnosis of PV modules and solar cells, there are two main approaches including electroluminescence and thermal imaging. In this study, the focus lies primarily on the first approach.

In the context of electroluminescence imaging, Bartler et al. [27] propose a deep learning-based classification pipeline consisting of: (1) an image pre-processing scheme for distortion correction, segmentation, and perspective correction; (2) a CNN model modified from the original VGG16 network architecture [28]. Additionally, they strongly address the general problem of data imbalance and overfitting in the automated classification of solar cell images and derive a combined method of non-heuristic oversampling [29] and data augmentation to ameliorate these shortcomings. Their results ascertain that with the combined method, the overfitting issue is reduced as indicated by a low balanced error rate (7.73%), and a low false negative rate (12.96%).

Deitsch et al. [30] propose two pipelines to determine the defect likelihood of an arbitrary solar cell. The first pipeline employs an SVM to discriminate various features extracted from captured electroluminescence images. These features are extracted using different methods. In the second pipeline, a CNN is suggested to discriminate between functional and defective cells. The proposed CNN is an adaption from the VGG19 network architecture [28], and it is trained with augmented data. Results have demonstrated that better classification is achieved using CNN (88.42%) than that SVM (82.44%).

Additionally, Tang et al. [31] propose to use CNN models for defect classification on an augmented electroluminescence image dataset. This dataset is a combination of traditional augmentation and image synthesis powered by GAN [32]. They attest that traditional augmentation is capable of generating new data with low computation time and hardware demand. Therefore, it is suitable to generate more data with simple image manipulation such as rotation, translation, and scaling. On the other hand, GAN-based image generation can potentially introduce more diversities to the current dataset. Hence, the combination of these will improve the classification accuracy significantly. Using the combined data augmentation approach and the proposed CNN model, their approach achieves significantly better results (83%) when compared with other pre-trained models such as VGG16 (66%), ResNet50 (67%) [33], and MobileNet (42%) [34].

Unsupervised Learning

Unsupervised learning - another classical field of machine learning, is a task of learning features from unlabeled data. One of the primary tasks of unsupervised learning is anomaly detection. Simply speaking, anomaly detection is the identification of infrequent observations that significantly deviate from the ‘normal behaviors’ of data. Thus, anomaly detection is essentially identical to fault detection. Interestingly, there are relatively few studies on how deep learning-based anomaly detection is realized in the context of defect detection in PV modules using electroluminescence images. As a result, many surveys in other areas have been conducted, including semiconductor device manufacturing and healthcare.

Carrera et al. [35] exploit the use of convolutional sparse models [36] in detecting anomalous regions in images. By learning a dictionary of filters of convolutional sparse representation for high-frequency components of images, anomalous regions can be detected at the patch level through an “outlier” indicator vector that falls outside a certain confidence region estimated from normal images. Besides successfully proving that convolutional sparse models perform much better than other standard patch-based sparsity methods, this study also points out the importance of local group sparsity in improving the performances of the models. Considering the spread of the non-zero coefficient across different maps results in a significant performance gap between different sparse models.

Davletshina et al. [37] propose to use of different unsupervised learning methods, including variations of GANs and auto-encoders, to detect anomalies in X-rays images of hands. They also introduce a powerful pipeline of preprocessing, consisting of cropping raw images using Otsu binarization [38], hand localization, fine-tuning, segmentation, and resizing before performing data augmentation and padding. By going through this powerful process, the effects of noise, which are usually mistaken to be an anomaly, are greatly reduced. Results have shown that all proposed models achieve their best runs with the highest ROC-AUC values (for instance, 60.7% for α-GAN and 57% for convolutional auto-encoder) when data is fully preprocessed, and most of them can generate highlighting of anomalous regions that even non-experts can interpret.

Shi et al. [39] propose a general unsupervised learning approach to detect and segment out small and confined anomalous regions in images. It features a pre-trained VGG19 [28] on ImageNet [40] to extract hierarchical image features, a multi-scale regional generator to generate discriminative multi-scale representations for every subregion of the image, and a deep convolutional auto-encoder to compress the multi-scale representation into a low-dimension latent space and reconstruct the representation again. The outcomes of this approach have proved that multi-scale representation is effective in detecting anomalies and the performance will increase as more and larger scales are leveraged, achieving the best ROC-AUC scores of 94.5% on object images and 93.3% on texture images.

Guo et al. [41] develop an approach based on one-class SVM [42] to detect and classify crystallographic defects in scanning transmission electron microscopy images [43]. They introduce 2 different methods of image segmentation and data preprocessing, depending on the number of domains in the images. Additionally, both methods use the Patterson function [44] as a feature extraction descriptor to greatly reduce the number of principal components, and thus, the variability in input space due to segmentation errors. Results have shown that this approach achieves satisfying results for defect detection in 2D materials and 3D nanocrystals, and can be applied to improve the crystal structure reconstruction of noisy atomic images.

3.3 Methods

Dataset

The proposed exploratory data analysis supervised learning, and unsupervised learning approaches are evaluated on the dataset that is publicly available [30, 45, 46], denoted as the ELPV dataset.

The ELPV dataset includes 2,624 samples of 300×300 pixels 8-bit grayscale images of functional (good) and defective (bad) solar cells with variable degrees of degradation. The samples were collected by segmenting and extracting individual solar cells from 44 different PV modules, 18 of which are monocrystalline and the rest are polycrystalline. Experts annotated each image in this dataset with a defect probability of 0.0 to 1.0 and the type of solar module (monocrystalline or polycrystalline) from which the solar cell image was retrieved. Table 1 provides the distribution of the total number of solar cells according to defect probability and wafer type.

Table 1: Distribution of data points in the dataset from Buerhop-Lutz and Deitsch et al. according to wafer type and defect probability.

Solar Wafer	**Labels**				**Total**
	0%	**33.33%**	**66.67%**	**100%**	
Monocrystalline	588	117	56	313	1,074
Polycrystalline	920	178	50	402	1,550
Total	1,508	295	106	705	2,624

The samples annotated with the defect probability of 33.33% and 66.67% are denoted as marginally defective. On the other hand, samples annotated with a probability of 100% are denoted as defective, and samples annotated with a 0% probability are denoted as functional. Since there are two wafer types (monocrystalline and polycrystalline), in total there are six groups consisting of mono-functional, mono-marginally-defective, mono-defective, poly-functional, poly-marginally-defective, and poly-defective.

Exploratory Data Analysis

The following holistic approach is proposed for exploratory data analysis to gain in-depth insight into the given dataset. Firstly, several statistical parameters, including mean, median, standard deviation, and mode, are calculated image-by-image. Considering that an image is a 2-D function of intensity $\mathrm{I(x,y)} = \mathrm{i_1(x,y), i_2(x,y), \ldots, i_j(x,y), \ldots, i_n(x,y)}$ where $\mathrm{i_j(x,y)}$ represents the intensity at a particular pixel and n is the total number of pixels in an image, the following methods are derived to compute these statistical parameters:

Mean of pixels $$\mathrm{mn} = \frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{i_j\,(x,y)}$$

Median of pixels $$\mathrm{md} = \frac{\mathrm{i}_{\mathrm{j}=\frac{\mathrm{n}}{2}}(\mathrm{x},\mathrm{y}) + \mathrm{i}_{\mathrm{j}=\frac{\mathrm{n}+1}{2}}(\mathrm{x},\mathrm{y})}{2}$$

Stdev of pixels $$\mathrm{std} = \sum_{\mathrm{j}=1}^{\mathrm{n}} \frac{\left(\mathrm{i}_{\mathrm{j}}(\mathrm{x},\mathrm{y}) - \mathrm{mn}\right)^2}{\mathrm{n}-1}$$

Maximum of pixels $$\mathrm{p}_{\mathrm{max}} = \mathrm{i}_{\mathrm{max}}(\mathrm{x},\mathrm{y}) \quad \forall \mathrm{i}_{\mathrm{j}}(\mathrm{x},\mathrm{y}) \in \mathrm{I}(\mathrm{x},\mathrm{y})$$

Minimum of pixels $$\mathrm{p}_{\mathrm{min}} = \mathrm{i}_{\mathrm{min}}(\mathrm{x},\mathrm{y}) \quad \forall \mathrm{i}_{\mathrm{j}}(\mathrm{x},\mathrm{y}) \in \mathrm{I}(\mathrm{x},\mathrm{y})$$

Mode of pixels $$\mathrm{p}_{\mathrm{mod}} = \mathrm{i}_{\mathrm{mod}}(\mathrm{x},\mathrm{y}) \quad \forall \mathrm{i}_{\mathrm{j}}(\mathrm{x},\mathrm{y}) \in \mathrm{I}(\mathrm{x},\mathrm{y})$$

Secondly, the average image is computed by taking the summation of all images associated with a specific group then dividing by the number of observations. Denoting N as the number of observations, k as an index of an observation $\mathrm{I}^{\mathrm{k}}(\mathrm{x},\mathrm{y})$ in a particular group, the method to compute an average image from samples of a specific group is derived as:

$$\bar{\mathrm{I}}(\mathrm{x},\mathrm{y}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{k}=1}^{\mathrm{N}} \mathrm{I}^{\mathrm{k}}(\mathrm{x},\mathrm{y})$$

Since the average images are representatives of their respective groups, they are then leveraged to compute the difference, also known as the contrast between good and bad samples. For instance, the contrast between good and bad samples of the monocrystalline wafer can be computed as follows:

$$\mathrm{C}_1(\mathrm{x},\mathrm{y}) = \bar{\mathrm{I}}_{\mathrm{functional}} - \bar{\mathrm{I}}_{\mathrm{marginally\ defective}}$$

$$\mathrm{C}_2(\mathrm{x},\mathrm{y}) = \bar{\mathrm{I}}_{\mathrm{functional}} - \bar{\mathrm{I}}_{\mathrm{defective}}$$

Finally, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are employed to visualize the distribution of the entire dataset.

Data Preparation

To prepare the dataset for this study, cell images annotated with a defect probability of 67% and of 33% (those that are classified as marginally defective) will be treated as fully defective. In other words, those images are annotated with a defect probability of 100%. Thus, the dataset will have 295 + 106 + 705 = 1106 defective samples and 1508 functional samples. The distribution of samples between these two classes is relatively even since the ratio between the number of functional samples and defective samples is approximately 1.36. Therefore, it can safely be assumed that issues related to data imbalance would unlikely occur.

After re-annotating the defect probability, the dataset is then partitioned into two 85:15 train versus test subsets, respectively. Next, the training subset is further split into two smaller 85:15 train versus validation subsets. As a result, there are a total of 1,895 training samples, 335 validation samples, and finally 394 test samples.

Additionally, data augmentation is introduced to generate slightly perturbed samples to expand this dataset artificially. The augmentation variability is kept modest because the segmented cells vary only by a few pixels along the translation axes. In this work, random flips are applied along the vertical and horizontal axes. Because the cell' busbar can be laid out either vertically or horizontally, the 90◦ random rotation is included. Additional methods used consist of random vertical and horizontal shifts and zooming that is limited to ±5% of the cell dimensions (height vs. width). Figure 1 illustrates the effects of the proposed data augmentation on the dataset.

Architecture Selection

To achieve the best possible results on the given dataset, the means of transfer learning and the "caviar" strategy are conducted. The "caviar" strategy refers to the practice of developing and training multiple models and selecting the model having the best learning curve. Following this strategy, a total of five models are proposed, four of which are supervised learning models and the final one is an unsupervised learning model.

Consider the supervised learning approach. Currently, there are more than 50 network architectures that deep learning practitioners can select for their intended classification tasks [47]. Consequently, it leads to the 'paradox of choice' where the availability of many options can problematize decision-making. To resolve this issue, the following approach, consisting of two steps for architecture selection, is proposed: First, only the models whose Top-1 accuracy is higher than 75% are considered to narrow down the number of options based on a common network scheme. From each scheme, the model with the highest Top-1 accuracy score will be picked. Using this approach, these four architectures have been chosen as follows:

- VGG19. This architecture is the best representative of the Stacking scheme, where convolution layers are stacked to get a deeper network [28].
- ResNet152v2. This architecture is the best candidate among those from the Residual scheme, where residual blocks in the network are employed to prevent the gradient vanishing problems [48].
- InceptionResNetv2. This architect has the best performance among networks that are built using the repetition of sparsely connected layers [49].
- NasNetLarge. This architecture is the best representative of the Neural
- Architecture Search scheme [50].

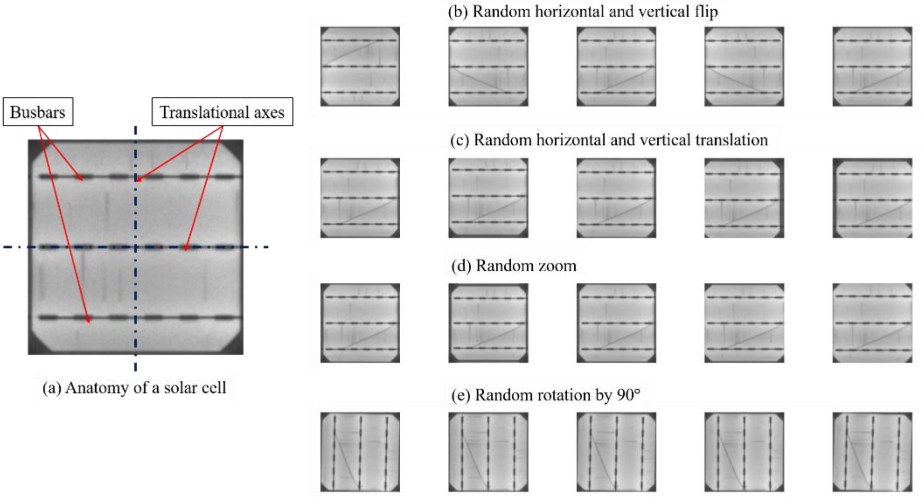

(b) Random horizontal and vertical flip
Busbars
Translational axes
(c) Random horizontal and vertical translation
(d) Random zoom
(a) Anatomy of a solar cell
(e) Random rotation by 90°

Figure 1: Effects of the proposed data augmentation on the dataset. (a) Anatomy of a solar wafer; (b) Results of random flips along the translation horizontal and vertical axes; (c) Results of random horizontal and vertical shifts; (d) Results of random zoom along the translation axes; (e) Results of random rotation by 90◦.

Having selected the suitable architectures, the four models, titled A, B, C, and D, are then implemented. Each model has the same number of fully connected layers (four of them), and the number of units for the first two fully connected layers is the same (4,096 units). The number of units for the third fully connected layer is selected to be equal to the number of kernels employed in the last convolutional layer in each respective model. ReLU is used as the activation function in all fully connected layers except for the final layer, and sigmoid is used as the activation function of the final fully connected layer because the output results are defined as defect probability. To avoid potential overfitting issues, two dropout layers are included, and the dropout rate is 20% for these layers. Figure 2 and Figure 3 present the architecture of the proposed models used for the prediction of defect probability in the electroluminescence images of solar cells.

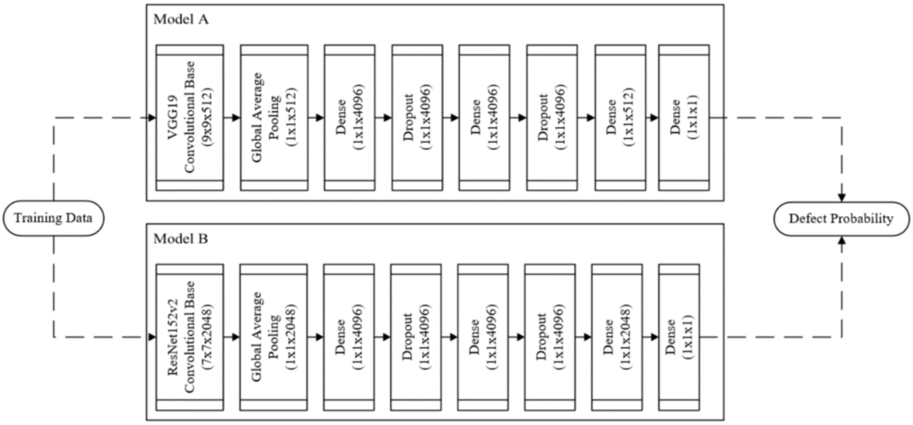


Figure 2: Architecture of the proposed classification models (Model A and Model B).

For the unsupervised learning approach, another model titled Model E is considered. The architecture of the proposed model is selected under the principles of a convolutional autoencoder, which is commonly employed for reconstruction-based anomaly detection. In this case, a VGG19 architecture with some modifications is used for the encoder component: the last three fully connected layers of the VGG19 architecture are stripped away, and only 19 convolutional layers are retained. For the decoder component, five transposed convolutional layers are used with a kernel size of three, and a stride of two. The number of filters employed in these layers is mirrored to those of the final layers in the convolutional blocks of VGG19, and the ELU function is selected as the activation function for these layers. The final layer of the decoder is a convolutional layer with three filters, a kernel size of three, and a linear activation function. Figure 4 presents the architecture of the proposed reconstruction model.

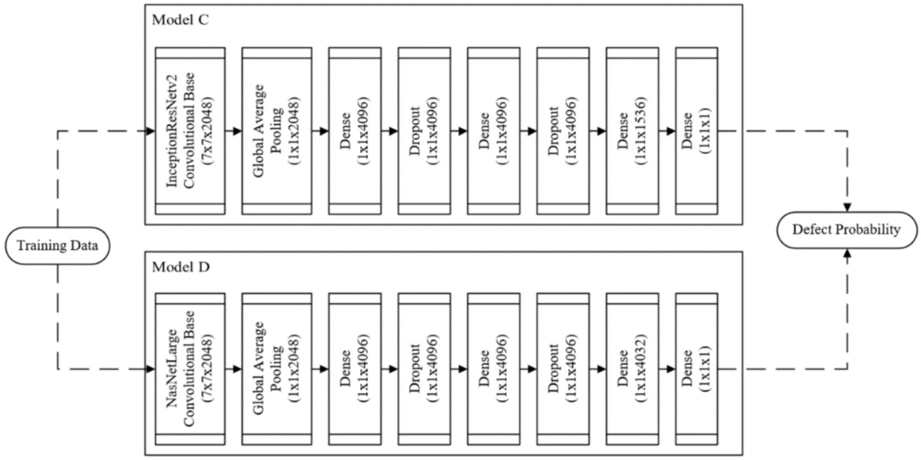


Figure 3: Architecture of the proposed classification models (Model C and Model D).

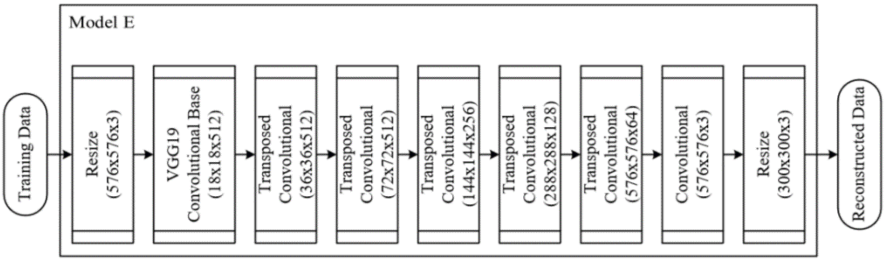

Model E
Training Data
Resize (576x576x3)
VGG19 Convolutional Base (18x18x512)
Transposed Convolutional (36x36x512)
Transposed Convolutional (72x72x512)
Transposed Convolutional (144x144x256)
Transposed Convolutional (288x288x128)
Transposed Convolutional (576x576x64)
Convolutional (576x576x3)
Resize (300x300x3)
Reconstructed Data

Figure 4: Architecture of the proposed reconstruction model (Model E).

Model Training

The training scheme for the proposed supervised and unsupervised learning models consists of two phases. First, the proposed models are loaded with the pre-trained ImageNet weights [40] before being trained, in a transfer-learning manner by only training the top layers (fully connected layers for classification models, transposed convolutional layers for the reconstruction model) with randomly initialized weights while keeping the weights of the convolutional bases immutable. In the second phase, the weights of all layers are refined.

For the first training phase, the Nadam optimizer [51] is deployed with a learning rate of $\eta = 10^{-3}$, exponential decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the regularization value $\epsilon = 10^{-8}$. For the second training phase, the Adam optimizer [52] is deployed with a smaller learning rate of $\eta = 10^{-5}$. The values of exponential decays β_1, β_2 and regularization value ϵ for Adam are identical to those of Nadam. Figure 5 illustrates the workflow of the proposed two-phase training scheme.

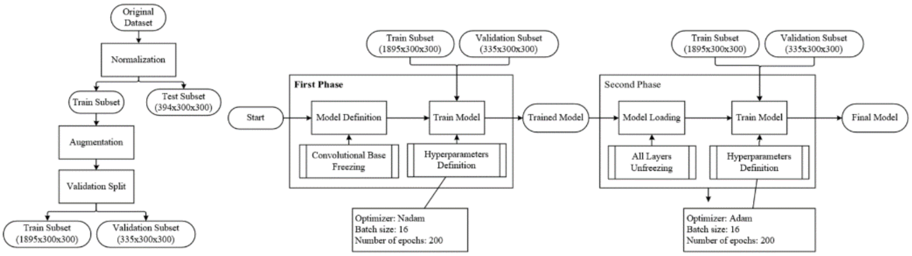


Figure 5: Training scheme for the proposed supervised and unsupervised learning models. The training consists of two phases: in the first phase, only the fully connected layers are trained; in the second phase, all layers are trained.

Since there are only two data classes (functional or 0.0 vs. defective or 1.0), binary cross-entropy loss is selected as the loss function for Models A, B, C, and D. Therefore, the training objective of the proposed classification models is the minimization of the cross-entropy empirical risk.

$$\mathrm{E}=\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{l}(\mathrm{y_i},\mathrm{p_i})$$

given:

$$\mathrm{l}(\mathrm{y_i},\mathrm{p_i})=\text{-}\mathrm{y_i}\log(\mathrm{p_i})\text{-}(1\text{-}\mathrm{y_i})\log(1\text{-}\mathrm{p_i})$$

with E is the empirical risk, n is the number of training samples, $l(y_i,p_i)$ is the cross-entropy loss function, y_i and p_i are ground-truth and estimated probability belonged to the positive class of a single training sample, respectively.

On the other hand, the purpose of the reconstruction method for anomaly detection is to measure the dissimilarity between the training data and the data reconstructed by a model given that (1) the model is solely trained on normal images, and (2) the trained model is unable to reproduce anomalous images [39]. Therefore, the structural dissimilarity metric loss is selected as the loss function for Model E since the training objective is to minimize the empirical pixel-wise distance between the training inputs and their respective reconstruction.

$$\mathrm{E}=\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{l}(\mathrm{y_i},\mathrm{p_i})$$

given:

$$\mathrm{l}(\mathrm{y_i},\hat{\mathrm{y}}_\mathrm{i})=\frac{1\text{-}\mathrm{SSIM}(\mathrm{y_i},\hat{\mathrm{y}}_\mathrm{i})}{2}$$

and:

$$\mathrm{SSIM}(\mathrm{y_i},\hat{\mathrm{y}}_\mathrm{i})=\frac{\left(2\mu_{\mathrm{y_i}}\mu_{\hat{\mathrm{y}}_\mathrm{i}}+\mathrm{c}_1\right)\left(2\sigma_{\mathrm{y_i}\hat{\mathrm{y}}_\mathrm{i}}+\mathrm{c}_2\right)}{\left({\mu^2}_{\mathrm{y_i}}+{\mu^2}_{\hat{\mathrm{y}}_\mathrm{i}}+\mathrm{c}_1\right)\left({\sigma^2}_{\mathrm{y_i}}+{\sigma^2}_{\hat{\mathrm{y}}_\mathrm{i}}+\mathrm{c}_2\right)}$$

with $\mathrm{SSIM}(\mathrm{y_i},\hat{\mathrm{y}}_\mathrm{i})$ is the structural similarity index; μ_{y_i} and $\mu_{\hat{y}_i}$ are pixel sample mean of ground-truth image $\mathrm{y_i}$ and its reconstruction counterpart $\hat{\mathrm{y}}_\mathrm{i}$, respectively; $\sigma_{\mathrm{y_i}\hat{y}_i}$ is the covariance between $\mathrm{y_i}$ and $\hat{\mathrm{y}}_\mathrm{i}$; ${\sigma^2}_{\mathrm{y_i}}$ and ${\sigma^2}_{\hat{y}_i}$ are the variances of $\mathrm{y_i}$ and $\hat{\mathrm{y}}_\mathrm{i}$; c_1 and c_2 are two variables to stabilize the division with weak denominator.

Each model is trained and validated on a single NVIDIA GeForce RTX 3080Ti, and the training session consists of a total of 200 epochs. The augmented versions of the training samples are processed in mini-batches of 16 samples and implement the proposed models using Keras version 2.9.0 [53] with TensorFlow version 2.9.1 [54]. When the training process is completed, the model with the best possible performance in terms of validation accuracy for each proposed architecture is saved.

Evaluation Metrics

Besides accuracy, other metrics are used to provide additional insights when measuring the performance of the models. These metrics are precision, recall, F1-score, underkill, overkill, AUC, and Matthews correlation coefficient (MCC, commonly known as ϕ coefficient). To qualitatively assess the model performance, two additional metrics are employed: class activation maps (CAMs) for supervised learning models, and heatmaps for unsupervised learning models.

Precision: Precision can also be understood as what proportion of identifications is correct. It is defined as the number of observations that are correctly classified as defective (true positive) over the total number of observations as defective by the model (total positive). Note that total positive includes true positive observations and those that are incorrectly classified as defective (false negative). As an example, a model with 80% precision when predicting a cell to be defective means that it is correct 80% of the time.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: The recall - commonly known as sensitivity - represents what proportion of actual defective observation was identified correctly. It is defined as the number of observations correctly classified as defective (true positive) over the total number of relevant observations. Note that the total number of relevant observations consists of true positive and false negative instances, and false negative is defined as the number of observations incorrectly classified as functional. Thus, a model with 55% recall means it correctly identifies 55% of all defective cells.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score: The F1-score is the harmonic mean of precision and recall metrics.

$$\text{F1-score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Underkill and **Overkill**: In defect inspection, underkill, or under rejection rate, is defined as the number of occurrences a system misses detecting defective products or items. Hence, underkill is computed as the number of false negative instances over the total number of predictions. In contrast, overkill is defined as the number of occurrences a system flags products or items as defective when they are not. Therefore, it is computed as the number of false positives over the total number of predictions.

$$\text{UR} = \frac{\text{FN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}$$

$$\mathrm{OR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}}$$

AUC: This metric is defined as the area under the receiver operating characteristic (ROC) curve, which represents the diagnostic ability of a classification system as its discrimination threshold is varied.

Matthews correlation coefficient: This coefficient essentially represents the correlation between predicted observations and their respective ground truths. It is considered to be one of the best measures of a classification system because it addresses well for even extremely imbalanced datasets.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

CAM and **Heatmap**: In supervised learning, CAM is the visualization of discriminative regions a CNN model uses to identify a specific class of a given image [55]. A high-performance CNN model generally has discriminative regions that are similar to those used by domain experts to classify an image to its corresponding class. In unsupervised learning, on the other hand, the heatmap is essentially a contrast image visualizing the difference between a given image and its corresponding reconstructed version.

Random Trial Experiment

Bengio et al. [56] state that pure randomness has a noticeable impact on deep learning models because their training involves several intrinsic randomness sources, including parameter initialization and example sampling. The initial state of the model parameters is also important due to the presence of local minima in the loss function. Hence, the choice of random seed during the train-test split and the results of the train-validation split can significantly affect the training results.

To evaluate the potential influence of random splitting on the dataset, a naïve random trial experiment is designed as follows: Firstly, the original dataset is shuffled, sampled, and partitioned into five seeds of 85:15 train and test subsets. Next, a model is selected from the four proposed classification models above. The chosen model is then trained (the first phase only, meaning the fully connected layers are trained) with different random seeds, each of which contains 1,895 training samples and 335 validation samples. For each seed, the model is trained 30 consecutive times. In each training time, validation metrics (accuracy, precision, and recall) belonging to an epoch that has the best validation accuracy are recorded. Using the recorded observations, a hypothesis test is conducted to quantify the impact of randomness on model training. Based on the proposed design of the experiment, there are in total of 150 observations for each metric of inter evenly divided for five seeds. Figure 6 and Figure 7 illustrate the design of the experiment for the random trial as well as the flowchart of a hypothesis test to gauge the targeted subjects, respectively.

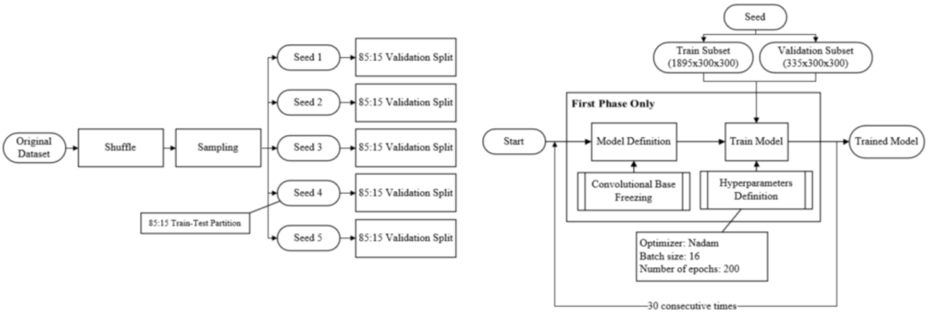


Figure 6: Design of experiment for the random trial experiment.

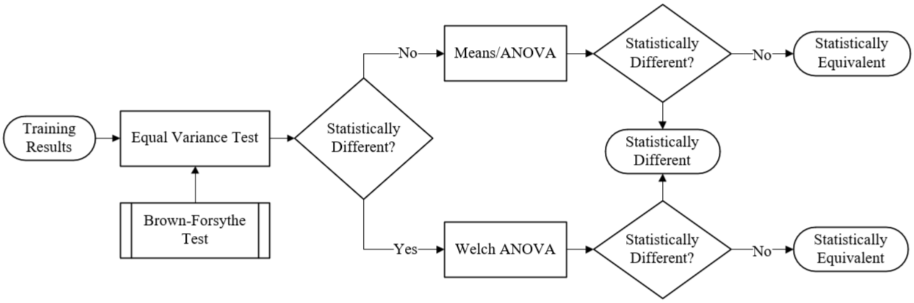


Figure 7: Procedure of the hypothesis test.

3.4 Results and Discussion

Exploratory Data Analysis

The distributions of statistical parameters for samples grouped by both cell types (monocrystalline vs. polycrystalline) and defect probabilities (0%: functional; 33% and 67%: marginally defective; 100%: defective) are visualized in Figure 8. It can be observed from Figure 8 that, the statistical parameters derived from defective samples (defective probability ≥ 33%) are typically lower compared to those of functional samples. In terms of mean, median, and mode of pixels, it is postulated that there is no statistical difference between the functional and marginally defective groups per cell type. Such validation test for the aforementioned claim, however, is unfeasible since there is a dis- parity between the sample size of these groups (1,508 functional images vs. 401 marginally defective images). Assuming that there is no statistical difference between these parameters, the proposed classification models are expected to have the less discriminative capability for the functional and marginally defective groups.

Figure 9 illustrates the average images derived from functional, marginally defective, and defective samples grouped by cell types. At first glance, there is no visual difference between these groups. The contrast images in Figure 10, however, demonstrate otherwise.

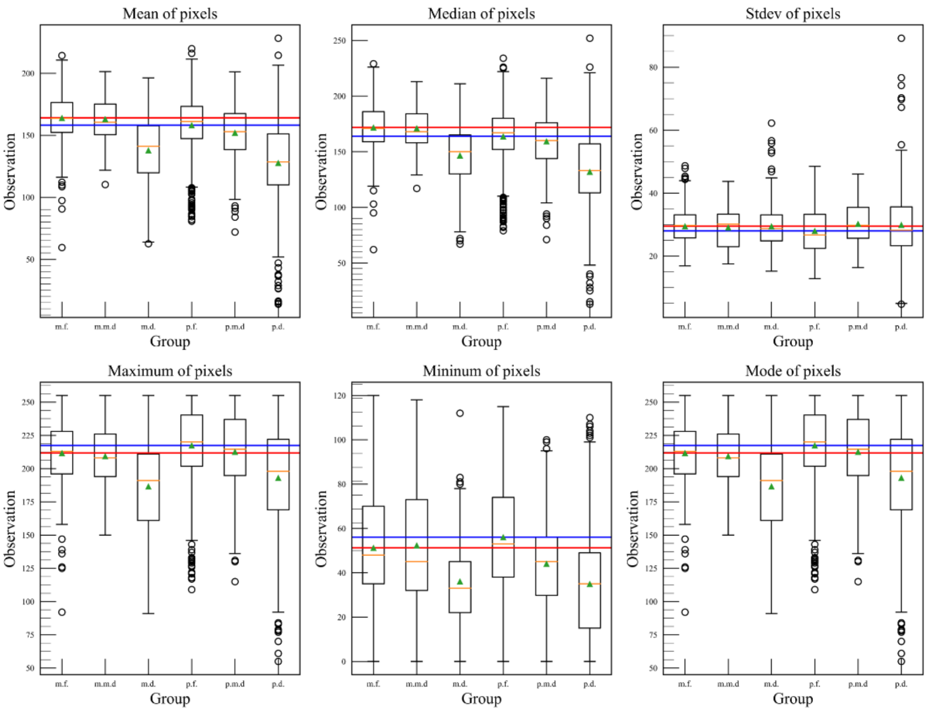


*Figure 8: Distributions of statistical parameters for samples grouped by both cell types and defect probability. Labels **m.f.**, **m.m.d.**, **m.d.**, **p.f.**, **p.m.d.**, **p.d.** are abbreviations for mono-functional, mono-marginally-defective, mono-defective, poly-functional, poly-marginally-defective, and poly-defective, respectively. The mean value of computed statistical parameters for mono-functional (red) and poly-functional (blue) are used as baselines.*

Figure 10 illustrates the dissimilarity between functional, marginally defective, and defective samples grouped by cell types. For monocrystalline samples, regions associated with the busbars display the most deviation. Thus, it can be theorized that defects occurring in monocrystalline solar cells are commonly materialized around the bus bars. On the other hand, most deviations between good and bad polycrystalline samples are located on both the left side and the middle of the cell. Hence, these regions can be postulated to be plagued with damages, deformities, and faults. Upon further verification [57], these conjectures proved to be valid for the following reasons:

- The cell's busbars usually experience disconnection failure mode since they act as a means for cell interconnection. To form an interconnection between two or more solar cells, a long metal conductor is soldered directly onto the busbars to form a bridge. Therefore, the reliability of the interconnection solely relies on the soldering quality.
- Micro-cracks (due to mechanical damage by various factors) and cell discoloration (mostly due to long-time usage) are commonly found on both sides and the middle area of the cell.

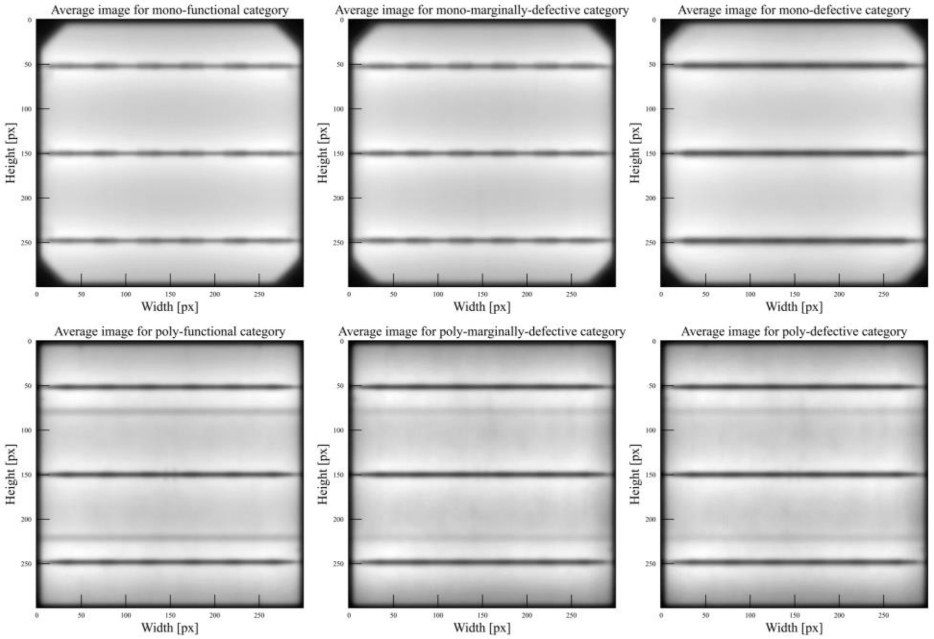


Figure 9: Computed average images between the functional, marginally defective, and defective samples.

The distributions of the dataset after embedding through PCA and t-SNE are shown in Figure 11. It can be seen that PCA fails to assemble the combined features, or principal components, from the dataset into sensible clusters. Therefore, it is possible that the combined features are not linear combinations of the dataset's original features. Moreover, it is possible that the preservation of large variance due to PCA's operating principles inadvertently minimizes the distance between embedding distributions of all sample groups. On the other hand, t-SNE manages to cluster the features into distinct groups. According to the results generated by t-SNE analysis, features associated with the mono-functional and mono-marginally-defective groups are usually clustered together. Features associated with the poly-functional and poly-marginally-defective groups also come hand-in-hand. Additionally, the t-SNE result draws better discrimination between a majority of samples from the functional and defective groups. However, there is no clear distinction between a significant quantity of features originating from the mono-defective and poly-functional groups.

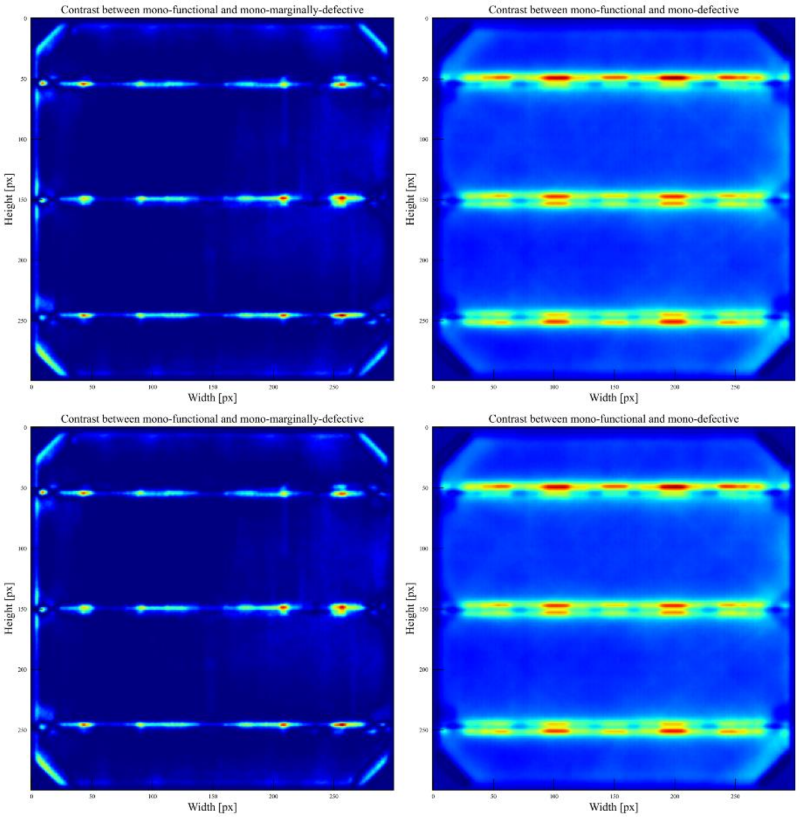
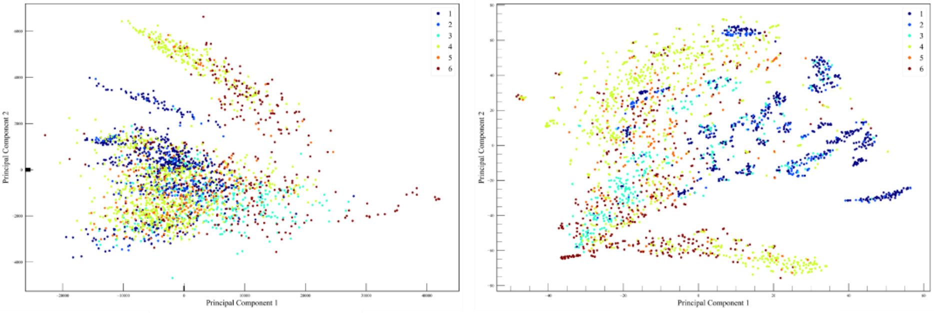


Figure 10: Computed contrast images between the functional, marginally defective, and defective samples. Regions that are most deviated are colored red, while most similar regions are colored blue.



(a) Embeddings using PCA

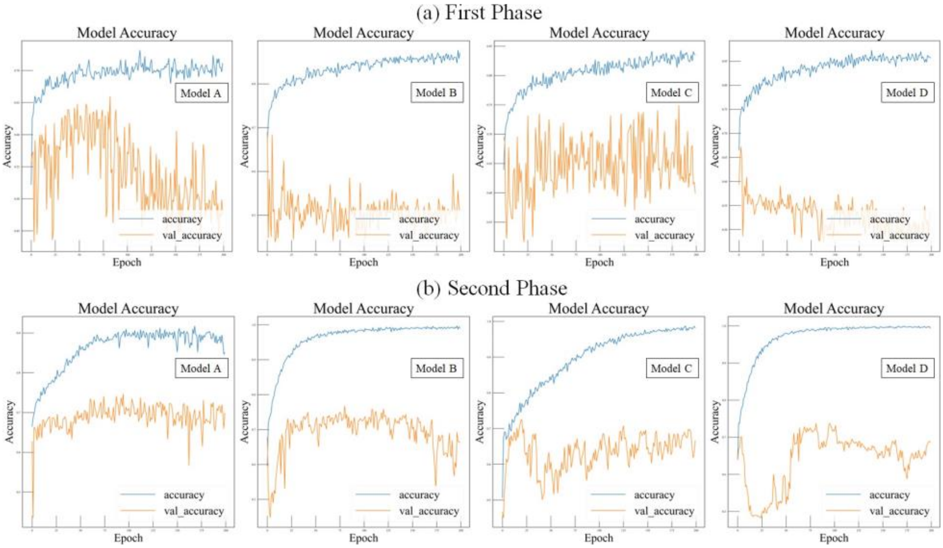
(b) Embeddings using t-SNE

Figure 11: Distribution of dataset embeddings: (a) Embedding using PCA; (b) Embedding using t-SNE. The sample classes are designated as follows: 1 stand for mono-functional; 2 mono-marginally-defective; 3 mono-defective; 4 poly-functional; 5 poly- marginally-defective; 6 poly-defective.

Model Performance

Figure 12 presents the accuracy performance of the training subset and validation subset in detail against the number of epochs during the training process for all proposed classification models. According to Figure 12, the accuracy for each respective model increases rapidly along with the increase in the number of epochs, and finally, the training process converges within around 100 epochs in both phases. Except for Model A, the proposed models achieve outstanding training accuracy results which are greater than 85%. However, all models experience a serious overfitting problem since there is a significant difference in mean classification accuracy between the train and validation subsets (≥ 10%). Such a problem is likely caused by the following reasons:

- The size of the training set is insufficient. As mentioned earlier, the ELPV dataset used in this study contains only 2,624 samples meaning there are only 1,895 samples for training given the train-test split is 85:15. Even with the proposed data augmentation scheme (random flip, random rotation, random shifts, and random zoom), the expected size of the training set is 1, 895 × 5 = 9, 475 samples, still smaller than the typical size of training sets employed in deep learning studies. Moreover, the results from the above PCA and t-SNE analyses indicate that more data are needed to increase models' discriminative capabilities.
- Model complexity. Since the proposed models contain three fully connected layers and each layer has a high number of units (≥ 1,000 units), the usage of regularizers including Lasso and Ridge as well as a higher dropout rate (e.g., 50%) are necessary to ameliorate the overfitting issue.

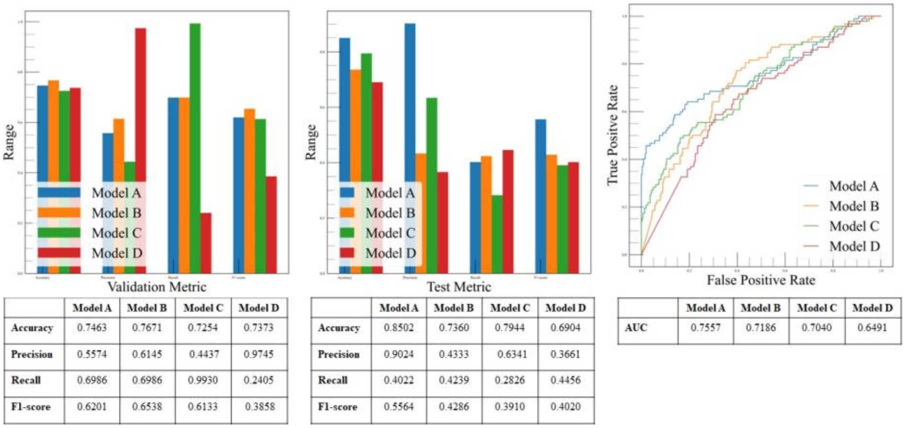

(a) First Phase

(b) Second Phase

Figure 12: Performance assessment of proposed classification models during the training process (first and second phases).

Figure 13 presents the accuracy, precision, recall, and F1-score results on the validation subset and one test set for all classification models. Figure 13 also provides the results of the ROC and AUC analysis for all models on the test set. The validation scores are extracted from an epoch containing the highest validation accuracy after the two-phase training process. It can be observed from Figure 13 that, Model B by far has the best overall performance since it possesses the highest validation accuracy (76.71%) and validation F1-score (65.38%). The second best-performed model is Model A, whose validation accuracy and validation F1-score are 74.63% and 62.01%, respectively. Interestingly, Model C and Model D, which are adopted from two models possessing some of the highest Top-1 accuracy scores, are outperformed by Model A and Model B in terms of validation accuracy and F1-score. Nevertheless, Model D has an exceedingly high precision rate (97.45%), and Model C has a near-perfect recall rate (99.30%). Hence, Model D is exemplary in identifying defects with high accuracy since the number of false positive cases is zero as the precision rate approaches 100%. Meanwhile, a near-perfect recall rate at Model C means that it can potentially detect most of the occurring defect cases.

When assessing the performance on the test subset, clearly Model A by far has the best performance since it has the highest test accuracy (85.02%), an outstanding test precision (90.24%), and the highest test F1-score (55.64%) among the proposed classification models. The AUC of Model A (75.57%) on the test set also agrees with the above statement: the best-performed model is Model A. The second-best model is Model B - whose test F1- score and AUC are 42.86% and 71.86%, respectively - followed by Model C and Model D.



	Model A	Model B	Model C	Model D
Accuracy	0.7463	0.7671	0.7254	0.7373
Precision	0.5574	0.6145	0.4437	0.9745
Recall	0.6986	0.6986	0.9930	0.2405
F1-score	0.6201	0.6538	0.6133	0.3858

	Model A	Model B	Model C	Model D
Accuracy	0.8502	0.7360	0.7944	0.6904
Precision	0.9024	0.4333	0.6341	0.3661
Recall	0.4022	0.4239	0.2826	0.4456
F1-score	0.5564	0.4286	0.3910	0.4020

	Model A	Model B	Model C	Model D
AUC	0.7557	0.7186	0.7040	0.6491

Figure 13: Performance assessment of proposed classification models on the validation subset and the test set.

Figure 14 shows the confusion matrices of the proposed classification models on the test subset. The threshold probability is designated to be 33.33%. Thus, when a cell image is inferred with a probability greater or equal to the threshold, its corresponding class is 1.0 or defective; otherwise, its class is 0.0 or functional. From Figure 14, the corresponding MCC scores, and underkill and overkill rates are calculated. These scores are summarized in Table 2. Model A by far is the most capable model since it procures the highest MCC score (56.48%), the lowest underkill rate (1.52%), and the second lowest overkill rate (12.69%). Concerning the MCC score, the second-best model is Model B (26.39%) followed by Model D (23.07%) and Model C (22.38%). Despite having the best overkill rate (5.84%), Model C has the worst underkill rate (37.31%).

Table 2: Corresponding MCC scores as well as underkill and overkill rates of the four proposed classification models in the same test subset.

	Model A	**Model B**	**Model C**	**Model D**
Underkill	0.0152	0.1371	0.3731	0.1802
Overkill	0.1269	0.1294	0.0584	0.1269
MCC	0.5648	0.2639	0.2238	0.23071

Based on the assessment results above, the most suitable classification model for the target of this study is Model A since it offers competitive accuracy scores, overkill, and underkill rates. The qualitative results in Figure 15 may provide clues on why Model A provides the best discriminative capability. It can be observed from Figure 15 that, Model A does provide the most sensible cues on why it infers a specific defect probability for a given observation: it heavily scrutinizes the side regions (top, down, left, and right sides) to discriminate whether a polycrystalline cell is defective. Such method is similar to the contrast image results highlighted in Figure 10. For monocrystalline samples, in contrast, Model A 'inspects' the entire cell area to infer a defect probability.

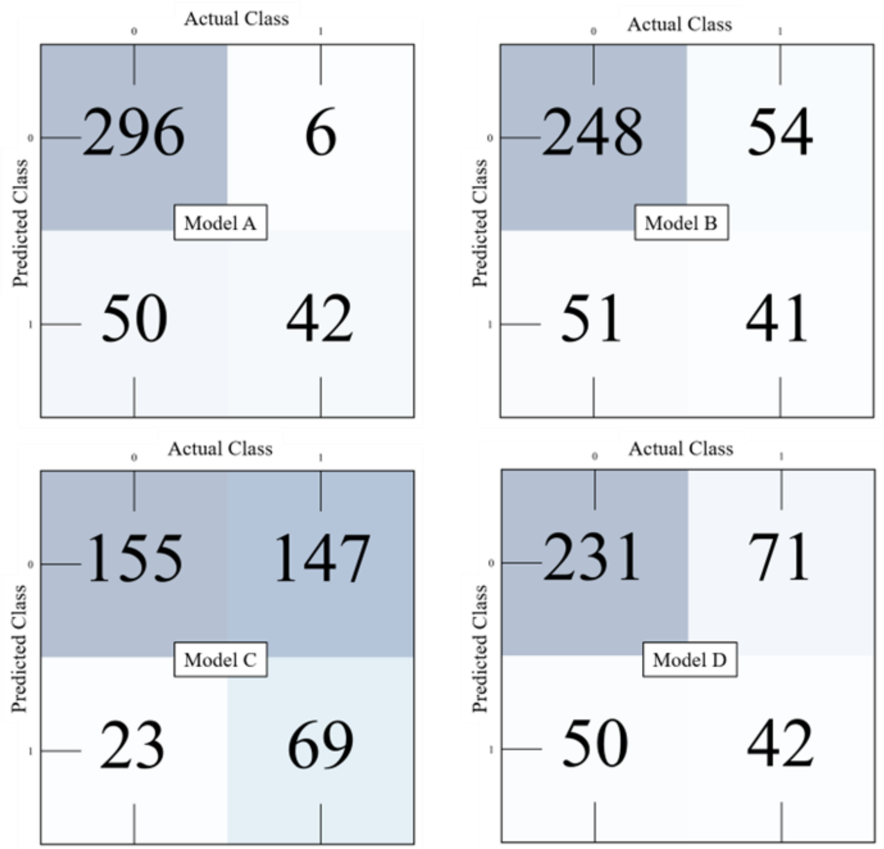


Figure 14: Confusion matrices of the four proposed classification models on the same test subset.

For the reconstruction model, Figure 16 presents the qualitative results of Model E on the test set. The discriminative threshold for the structural dissimilarity metric score, which is employed to infer whether a cell image is defective, is selected to be 0.33333333. According to Figure 16, Model E manages to reconstruct images containing solar cells with good accuracy. When Model E is used to infer cell images containing defectives such as micro-cracks or darken cells, it fails to discriminate correctly. Moreover, the generated heatmaps fail to pinpoint the locations of defects and faults in the cell. Hence, the reconstruction-based approach for fault detection with the given dataset is unfeasible.

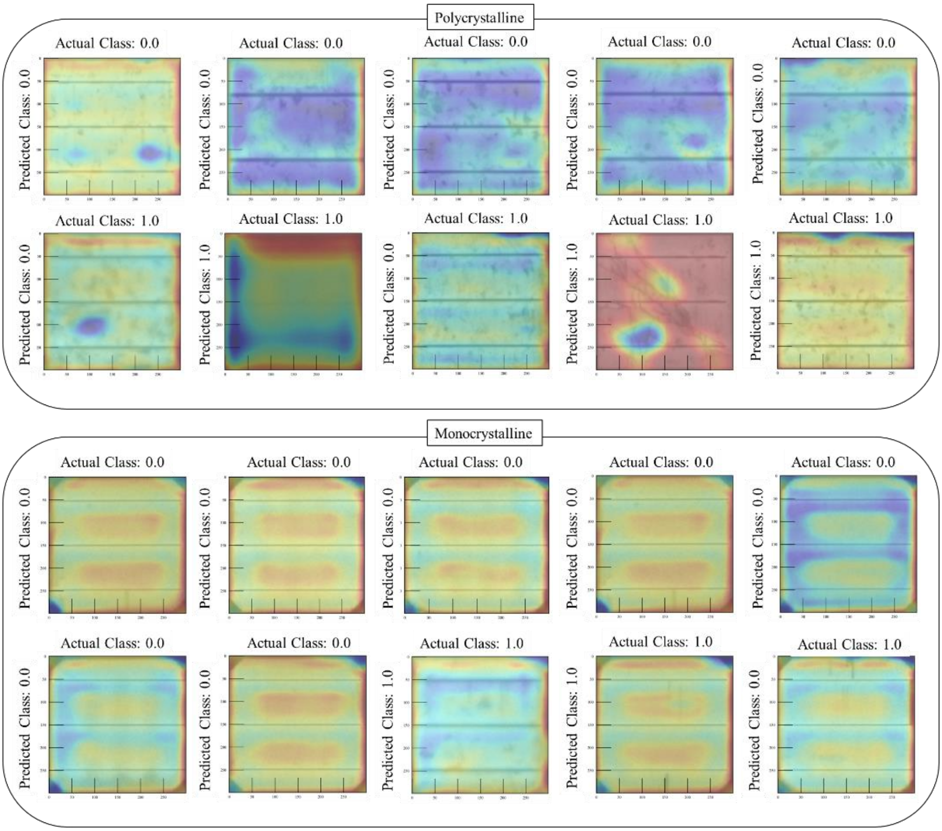


Figure 15: Selected qualitative results of Model A on the test subset. Red regions are areas where the model heavily uses to infer a specific probability.

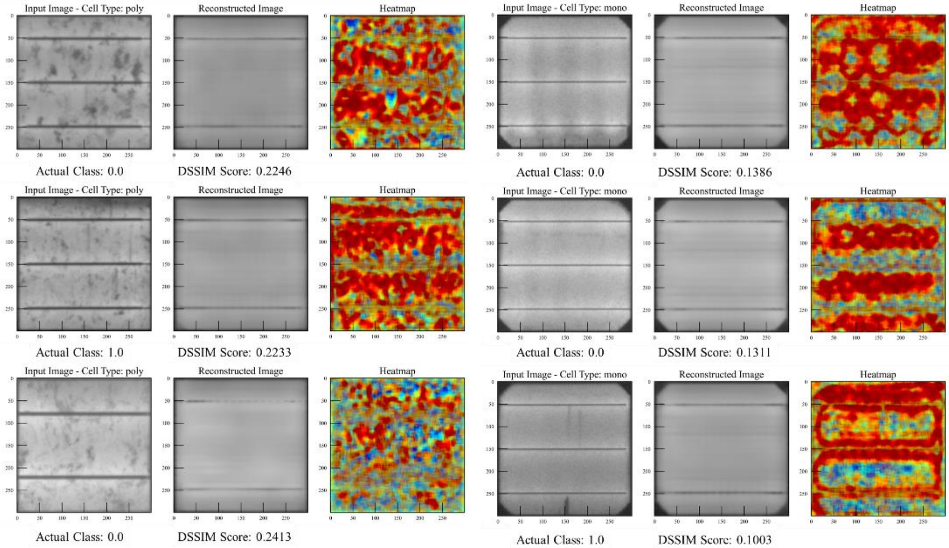


Figure 16: Selected qualitative results of Model E on the test subset. Red regions are areas having the most pixel-wise deviation between the input image and its respective reconstruction image.

Random Trial Experiment

Figure 17 shows the distribution of recorded validation metrics including accuracy, precision, and recall in the random trial experiments, and Table 3 shows the results of the equal variance and ANOVA hypothesis tests. Based on the model assessment results, Model A is selected to be the target of this experiment since it exhibits the best performance in most selection metrics. According to Figure 17, there are disparities in model performance when it is trained with different seeds of training data.

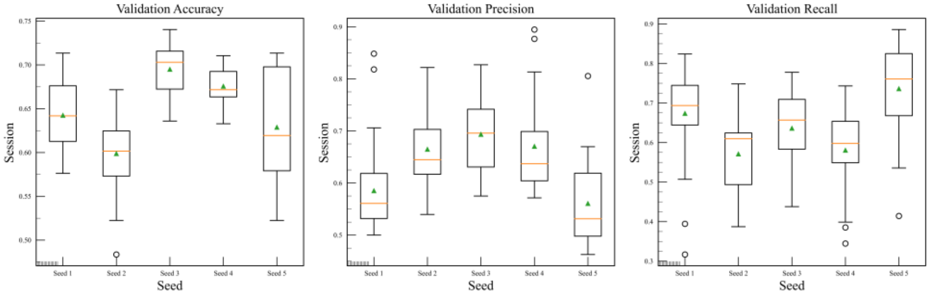


Figure 17: Distribution of recorded validation metrics (accuracy, precision, and recall) in the random trial experiment.

From Table 3, the results of the Brown-Forsythe test demonstrate that except for validation accuracy, other recorded metrics are likely to have the same variance since the null hypothesis fails to be rejected (p-value is greater than 0.05). Hence, it can be concluded that random splits of training data into train and validation subsets have a significant impact on the model's performance. Thus, the values of precision and recall will be subjected to the one-way ANOVA test, while Welch's ANOVA test will be conducted for values of accuracy. Results of the ANOVA tests clearly show that the null hypothesis, i.e., there is no difference between five seeds, is rejected. Therefore, disparities in model performance due to different seeds of training data and different validation splits are statistically significant.

Table 3: Results of the equal variance and analysis of variance (ANOVA) hypothesis tests on the recorded metrics in the random trial experiment. Noted that the accuracy is subjected to Welch's ANOVA test.

Brown-Forsythe Test (Equal variances)	$\mathrm{H_o}: \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5$ $\mathrm{H_a}: \sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4 \neq \sigma_5$	
	Accuracy	$F - test\ =\ 8.2583$ $p - value = 4.986e^{-6}$
	Precision	$F - test\ =\ 0.1304$ $p - value = 0.9710$
	Recall	$F - test\ =\ 0.2067$ $p - value = 0.9343$
ANOVA Test (Equal means)	$\mathrm{H_o}: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ $\mathrm{H_a}$: at least one μ_k is different	
	Accuracy*	$F - test\ =\ 30.7148$ $p - value = 3.500e^{-6}$
	Precision	$F - test\ =\ 15.4489$ $p - value = 1.506e^{-10}$
	Recall	$F - test\ =\ 13.374$ $p - value = 2.66e^{-9}$

Since there are disparities in model performance post-train due to the choice of random seed and random train-validation split, it is recommended to employ the K-fold cross-validation, especially the leave-one-out cross-validation [58] since the dataset used in this study is appropriately small, so that the model can learn better representations. In addition, the number of epochs should be sufficiently large to cover all possible permutations of random validation splits. Such number of epochs can be estimated as follows:

$$N = \frac{n \times D \times d}{b}$$

With N is the number of epochs, n is the number of training samples, D is the number of data augmentation, d is the amount of augmentation a single training observation is expected to have, and b is the batch size. In this case, n is 1,895, D is 4 since there are four types of employed data augmentation (random flip, random shift, random zoom, and random rotation), and d is 16 as there are $2^4 = 16$ combinations of augmentation types, and b is 16 as the batch size is defined to be 16. Therefore, the number of epochs is 1,895.

Further Research

According to the assessment results, Model A has the best performance among the proposed classification models given the above-defined hyperparameters (learning rates, optimizers, number of epochs, etc.) and training scheme. This raises a question on how optimal hyperparameters can be searched for to yield the best model out of the proposed Model A. Besides, assuming that different versions of Model A with different sets of hyperparameters are initialized, there is also a consideration of how these models can be gauged and compared to select the best one. In further research, hyperparameter optimization based on random search as well as the hypothesis test proposed by McNemar et al. [59] will be investigated to see how they are applicable to this study.

On the other hand, assessment results of Model E demonstrate that the reconstruction-based method is unlikely feasible. It, however, is possible to improve the reconstruction capability by employing a different loss function or architecture. In further research, perceptual loss [60] and uNet architecture [61] will be studied to see whether they can enable the application of the reconstruction-based method for this study.

3.5 Code Availability

Code for data cleaning and analysis is provided as part of the replication package. It is available here.

References

[1] Bosman, L.B., Leon-Salas, W.D., Hutzel, W., Soto, E.A.: PV system predictive maintenance: Challenges, current approaches, and opportunities. Energies 13(6) (2020). https://doi.org/10.3390/en13061398

[2] Rangaraju, S., Isaac, O., Vo, P., Nguyen, K., Ananth, A.: Guaranteed O&M for solar plants in Vietnam - a review proposal on guaranteed O&M service foster sustainable energy generation by maximizing solar energy production and safeguarding investment. International Journal of Engineering Applied Sciences (IJEAS) 8, 24 (2021). https://doi.org/10. 31873/IJEAS.8.7.08

[3] Rangaraju, S., Isaac, O., Vo, P., Ghosh, A., Kumaravel, S.: Robotizing solar farms in Vietnam - an optimal solution to reduce maintenance cost and increase efficiency. (2022). https://doi.org/10.6084/m9.figshare. 19543843.v1

[4] Aithagga, A., Assmus, J., Aubagnac, R., Auger, G., Barandalla, D., Bar- tle, M., Beauvais, A., Beggi, A., Bernardi, E., Berry, M., et al.: Operation & Maintenance Best Practices Guidelines.

[5] Shin, W., Han, J., Rhee, W.: AI-assistance for predictive maintenance of renewable energy systems. Energy 221, 119775 (2021)

[6] Betti, A., Trovato, M.L.L., Leonardi, F.S., Leotta, G., Ruffini, F., Lanzetta, C.: Predictive maintenance in photovoltaic plants with a big data approach. arXiv preprint arXiv:1901.10855 (2019)

[7] Nabti, M., Bybi, A., Garoum, M., et al.: Machine learning for predictive maintenance of photovoltaic panels: cleaning process application. In: E3S Web of Conferences, vol. 336, p. 00021 (2022). EDP Sciences

[8] Huuhtanen, T., Jung, A.: Predictive maintenance of photovoltaic panels via deep learning. In: 2018 IEEE Data Science Workshop (DSW), pp. 66–70 (2018). IEEE

[9] Gligor, A., Dumitru, C.-D., Grif, H.-S.: Artificial intelligence solution for managing a photovoltaic energy production unit. Procedia Manufacturing 22, 626–633 (2018)

[10] Kalogirou, S., Sencan, A.: Artificial intelligence techniques in solar energy applications. Solar Collectors and Panels, Theory and Applications 15, 315–340 (2010)

[11] Milidonis, K., Blanco, M.J., Grigoriev, V., Panagiotou, C.F., Bonanos, A.M., Constantinou, M., Pye, J., Asselineau, C.-A.: Review of applications of AI techniques to solar tower systems. Solar Energy 224, 500–515 (2021)

[12] Detollenaere, A., Van Wetter, J., Masson, G., Kaizuka, I., J¨ager-Waldau, A., Donoso, J.: Snapshot of Global PV Markets 2020 PVPS Task 1 Strategic PV Analysis and Outreach (2020). https://doi.org/10.13140/RG.2.2. 24096.74248

[13] Masson, G., Bosch, E., Kaizuka, I., Jäger-Waldau, A., Donoso, J.: Snapshot of Global PV Markets 2022 Task 1 Strategic PV Analysis and Outreach PVPS, (2022)

[14] Glavaski, S., Elgersma, M.: Active aircraft fault detection and isolation. In: 2001 IEEE Autotestcon Proceedings. IEEE Systems Readiness Technology Conference. (Cat. No.01CH37237), pp. 692–705 (2001). https://doi.org/10.1109/AUTEST.2001.949453

[15] Arnanz, R., Santiago, M., Dom´ınguez, A., Rodr´ıguez, J., Saludes-Rodil, S.: Monitoring and Fault Diagnosis in Manufacturing Processes in the Automotive Industry, (2011). https://doi.org/10.5772/13307

[16] López, B., Mel´endez, J., Wissel, H., Haase, H., Laatz, K.: Towards medical device maintenance workflow monitoring (2009)

[17] Chien, C.-F., Hsu, C.-Y., Chen, P.-N.: Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence. Flexible Services and Manufacturing Journal 25 (2012). https://doi.org/10. 1007/s10696-012-9161-4

[18] Ignacio Torrens, J., Keane, M., Costa, A., O'Donnell, J.: Multi-criteria optimisation using past, real-time and predictive performance benchmarks. Simulation Modelling Practice and Theory 19(4), 1258–1265 (2011). https://doi.org/10.1016/j.simpat.2010.11.002. Sustainable Energy and Environmental Protection "SEEP2009"

[19] Mellit, A., Tina, G.M., Kalogirou, S.A.: Fault detection and diagnosis methods for photovoltaic systems: A review. Renewable and Sustainable Energy Reviews 91, 1–17 (2018). https://doi.org/10.1016/j.rser.2018.03.062

[20] Isermann, R.: Model-based fault-detection and diagnosis – status and applications. Annual Reviews in Control 29(1), 71–85 (2005). https://doi.org/10.1016/j.arcontrol.2004.12.002

[21] Katipamula, S., Brambley, M.: Methods for fault detection, diagnostics and prognostics for building systems - a review part I. HVAC and R Research 11 (2005). https://doi.org/10.1080/10789669.2005.10391133

[22] Katipamula, S., Brambley, M.: Review article: Methods for fault detection, diagnostics, and prognostics for building systems—a review, part II. HVACR Research 11, 3–25 (2005). https://doi.org/10.1080/10789669.2005.10391123

[23] Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., Gao, R.: Deep learning and its applications to machine health monitoring: A survey (2016)

[24] Heckert, N.A., Filliben, J.J., Croarkin, C.M., Hembree, B., Guthrie, W.F., Tobias, P., Prinz, J., et al.: Handbook 151: NIST/Sematech e-handbook of statistical methods (2002)

[25] Engan, K., Eftestøl, T., Ørn, S., Kvaløy, J.T., Woie, L.: Exploratory data analysis of image texture and statistical features on myocardium and infarction areas in cardiac magnetic resonance images. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 5728–5731 (2010). https://doi.org/10.1109/IEMBS.2010.5627866

[26] Brewka, G.: Artificial intelligence - a modern approach by Stuart Russell and Peter Norvig, Prentice Hall. Series in artificial intelligence, Englewood Cliffs, NJ. The Knowledge Engineering Review 11(1), 78–79 (1996)

[27] Bartler, A., Mauch, L., Yang, B., Reuter, M., Stoicescu, L.: Automated detection of solar cell defects with deep learning. In: 2018 26th European Signal Processing Conference (EUSIPCO), pp. 2035–2039 (2018). https://doi.org/10.23919/EUSIPCO.2018.8553025

[28] Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv (2014). https://doi.org/10.48550/ ARXIV.1409.1556. https://arxiv.org/abs/1409.1556

[29] Garcıa, V., Sanchez, J.S., Mollineda, R.A.: Exploring the performance of resampling strategies for the class imbalance problem. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 541–549 (2010). Springer

[30] Deitsch, S., Christlein, V., Berger, S., Buerhop-Lutz, C., Maier, A., Gall-Witz, F., Riess, C.: Automatic classification of defective photovoltaic module cells in electroluminescence images. Solar Energy 185, 455–468 (2019). https://doi.org/10.1016/j.solener.2019.02.067

[31] Tang, W., Yang, Q., Xiong, K., Yan, W.: Deep learning based automatic defect identification of photovoltaic module using electroluminescence images. Solar Energy 201, 453–460 (2020). https://doi.org/10.1016/j. solener.2020.03.049

[32] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, Y.: A. and bengio. generative adversarial nets. In: Proceedings Neural Information Processing Systems, pp. 2672–2680

[33] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv (2015). https://doi.org/10.48550/ARXIV.1512.03385. https://arxiv.org/abs/1512.03385

[34] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv (2017). https://doi.org/ 10.48550/ARXIV.1704.04861. https://arxiv.org/abs/1704.04861

[35] Carrera, D., Boracchi, G., Foi, A., Wohlberg, B.: Detecting anomalous structures by convolutional sparse models. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2015). https://doi. org/10.1109/IJCNN.2015.7280790

[36] Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional net- works. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2528–2535 (2010). https://doi.org/10.1109/ CVPR.2010.5539957

[37] Davletshina, D., Melnychuk, V., Tran, V., Singla, H., Berrendorf, M., Faerman, E., Fromm, M., Schubert, M.: Unsupervised Anomaly Detection for X-Ray Images. arXiv (2020). https://doi.org/10.48550/ARXIV.2001. 10883

[38] Yousefi, J.: Image Binarization Using Otsu Thresholding Algorithm. https://doi.org/10.13140/RG.2.1.4758.9284

[39] Shi, Y., Yang, J., Qi, Z.: Unsupervised anomaly segmentation via deep feature reconstruction. Neurocomputing 424, 9–22 (2021). https://doi. org/10.1016/j.neucom.2020.11.018

[40] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Ima- geNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115(3), 211–252 (2015). https://doi.org/10. 1007/s11263-015-0816-y

[41] Guo, Y., Kalinin, S.V., Cai, H., Xiao, K., Krylyuk, S., Davydov, A.V., Guo, Q., Lupini, A.R.: Defect detection in atomic-resolution images via unsupervised learning with translational invariance. NPJ Computational Materials 7(1), 180 (2021). https://doi.org/10.1038/s41524-021-00642-1

[42] Scholkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. In: Solla, S., Leen, T., Mu¨ller, K. (eds.) Advances in Neural Information Processing Systems, vol.12. MIT Press, (1999). https://proceedings.neurips.cc/paper/1999/ file/8725fb777f25776ffa9076e44fcfd776-Paper.pdf

[43] Ziatdinov, M., Ghosh, A., Wong, T., Kalinin, S.V.: Atomai: A deep learning framework for analysis of image and spectroscopy data in (scanning) transmission electron microscopy and beyond. arXiv preprint arXiv:2105.07485 (2021)

[44] Patterson, A.L.: A Fourier series method for the determination of the components of interatomic distances in crystals. Phys. Rev. 46, 372–376 (1934). https://doi.org/10.1103/PhysRev.46.372

[45] Buerhop-Lutz, C., Deitsch, S., Maier, A., Gallwitz, F., Berger, S., Doll, B., Hauch, J., Camus, C., Brabec, C.J.: A benchmark for visual identification of defective solar cells in electroluminescence imagery. In: European PV Solar Energy Conference and Exhibition (EU PVSEC) (2018). https:// doi.org/10.4229/35thEUPVSEC20182018-5CV.3.15

[46] Deitsch, S., Buerhop-Lutz, C., Sovetkin, E., Steland, A., Maier, A., Gallwitz, F., Riess, C.: Segmentation of photovoltaic module cells in uncalibrated electroluminescence images 32(4). https://doi.org/10.1007/ s00138-021-01191-9

[47] Bianco, S., Cad`ene, R., Celona, L., Napoletano, P.: Benchmark analysis of representative deep neural network architectures. IEEE Access 6, 64270– 64277 (2018). https://doi.org/10.1109/ACCESS.2018.2877890

[48] He, K., Zhang, X., Ren, S., Sun, J.: Identity Mappings in Deep Residual Networks. arXiv (2016). https://doi.org/10.48550/ARXIV.1603.05027. https://arxiv.org/abs/1603.05027

[49] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, Inception- ResNet and the Impact of Residual Connections on Learning. arXiv (2016). https://doi.org/10.48550/ARXIV.1602.07261. https://arxiv.org/ abs/1602.07261

[50] Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning Transferable Architectures for Scalable Image Recognition. arXiv (2017). https://doi. org/10.48550/ARXIV.1707.07012. https://arxiv.org/abs/1707.07012

[51] Dozat, T.: Incorporating Nesterov momentum into Adam. (2016)

[52] Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv (2014). https://doi.org/10.48550/ARXIV.1412.6980. https://arxiv. org/abs/1412.6980

[53] Chollet, F., et al.: Keras. https://keras.io (2015)

[54] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Man´e, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Vi´egas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org (2015). https://www.tensorflow.org/

[55] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. arXiv (2015). https://doi. org/10.48550/ARXIV.1512.04150. https://arxiv.org/abs/1512.04150

[56] Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. Springer (2012)

[57] K¨ontges, M., Kurtz, S.R., Packard, C.E., Jahn, U., Berger, K.A., Kato, K., Friesen, T., Liu, H., van Iseghem, M., Wohlgemuth, J.H., Miller, D.C., Kempe, M.D., Hacke, P., Reil, F., Bogdanski, N., Herrmann, W., Buerhop- Lutz, C., Razongles, G., Friesen, G.: Review of failures of photovoltaic modules. (2014)

[58] Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction vol. 2. Springer, (2009)

[59] Thomas G. Dietterich; Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput 1998; 10 (7): 1895–1923. doi: https://doi.org/10.1162/089976698300017197

[60] Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. arXiv (2016). https://doi.org/10.48550/ ARXIV.1603.08155. https://arxiv.org/abs/1603.08155

[61] Feng, J., Deng, J., Li, Z., Sun, Z., Dou, H., Jia, K.: End-to-end res-unet based reconstruction algorithm for photoacoustic imaging. Biomed. Opt. Express 11(9), 5321–5340 (2020). https://doi.org/10.1364/BOE.396598