Received: 30 June 2021 | Revised: 23 August 2021 | Accepted: 23 August 2021

DOI: 10.1002/pip.3469

SPECIAL ISSUE ARTICLE

PROGRESS IN PHOTOVOLTAICS WILEY

Advanced analytics on IV curves and electroluminescence images of photovoltaic modules using machine learning algorithms

Vedant Kumar | Pranav Maheshwari

PV Diagnostics, Fototentia Diagnostics Private Limited, Mumbai, Maharashtra, India

Correspondence
Pranav Maheshwari, PV Diagnostics, Fototentia Diagnostics Private Limited, B/702, Satellite Garden 1, Gen AKV Road, Goregaon East, Mumbai, Maharashtra 400063, India.
Email: pranav@pv-diagnostics.com

Funding information
PV Diagnostics

Abstract

Advanced analysis and monitoring of photovoltaic solar modules is required to maintain the reliable operations of photovoltaic plants. Hence, it requires diagnostics through current–voltage (IV) curves, electroluminescence (EL) imaging, and other measurement techniques. The analysis through IV characterization provides the discerning insight about the quantitative measure of solar module performance, while the image characterization methods on EL images can capture spatial defects with microscopic resolution such as microcracks, broken cells interconnections, shunts, among many other defect types. The fusion of these two methods with supervised and unsupervised machine learning can generate unique insight with classification, regression, and dimension reductions on IV–EL data. In this study, we have performed the IV–EL correlation by classifying the IV data based on EL image annotation (where the class information is coming from EL image). The feature vectors consist of IV curve parameters and statistical features. We have first applied the unsupervised learning algorithms *t*-distributed stochastic neighbor embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP) for dimensionality reduction to understand the importance of various features on EL defect types. Furthermore, we had applied feature selection algorithms before applying the classification algorithms. We have performed the classification of various defect types by applying the random forests (RF) and XGBoost algorithm while identifying the top features. The accuracy was achieved greater than 91% and 95%, respectively, for supervised methods on the top five features. This correlation of IV–EL measurement could benefit in quick identification of various defect types in PV modules with only IV curve parameters, given the classification models are modeled using large-scale datasets and tuned optimally.

KEYWORDS
classification, correlation, dimension reduction, electroluminescence imaging, IV curve, random forests, *t*-SNE, UMAP, XGBoost

1 | INTRODUCTION

The global installation of large photovoltaic (PV) plants has grown significantly in the last decade. The global installed solar PV capacity exceeded 500 GW at the end of 2018, and an estimated additional 500 GW of PV capacity is projected to be installed by 2022–2023, bringing us into the era of TW-scale PV.[1] It is proven that PV technology can only be economical if PV modules achieve reliability and high performance for 25–30 years under field conditions.[2] The installation and operations of PV modules undergo complex processes that often lead to degradation due to various exposures to environmental conditions as well as mechanical stress during transportation.[3] The most common stressor from environmental conditions are temperature, humidity, wind, soiling, and ultraviolet (UV) irradiance, which leads to the emergence of defective behavior through degradation modes that include cell cracking, series resistance increases, discoloration, hot spots, encapsulant chemical changes, and many others.[4,5]

To maintain the PV plant operation and power generation at the maximum capacity, it is obligatory to perform the diagnostic of PV modules on recurrent basis. Identifying the defects and faults in PV modules rapidly is a challenging task, yet it can benefit the solar power generation industry exceedingly. There are numerous measurement techniques available for PV modules test and diagnostics. One of the most effective diagnostics techniques to assess the module degradation and performance are maximum power point tracking and current–voltage (IV) curve tracing. The IV curve tracing provides point-in-time data regarding the electrical state of the test module.[4] Of these, IV tracing performs a complete electrical sweep from short-circuit current (I_{SC}) to open-circuit voltage (V_{OC}) of an illuminated PV module. The decreasing short-circuit current (I_{SC}) and shunt resistance (R_{SH}) are sometimes related to cracked cells,[6] and increasing series resistance (R_S) can be related to corrosion of cell metallization or interconnects. But IV curve features alone cannot provide the insights regarding degradation mechanisms leading to power loss.

The visual images of PV modules could be useful as well in assessing the performance and degradation behavior, but these studies are usually qualitative and observational in nature. There are range of different other imaging techniques such as electroluminescence (EL), photoluminescence, UV fluorescence, and IR thermography, which can be utilized to collect high density quantitative data that provide information on cell variability and other module defects.[7] The EL imaging is one such technique, which helps in identifying PV module defects such as cell cracks, potential-induced degradation (PID), crystal dislocations in a multicrystalline wafer, finger interruptions, humidity corrosion, shunted bypass diode (substring failure), and many others. This is possible because EL images are rich in spatial information.[4]

There is a scarcity of research of correlation studies of IV data and EL imaging of PV modules. Many insights can be drawn from IV curve and EL image correlation, which can lead to identification of various defect types. There are few research articles in similar domain, but experiments in those studies have been performed in controlled environment.[8] In this research, we have devised a unique machine learning approach for identifying various defect types through correlation of IV curve data and EL images, where we have trained our classification model on features generated from IV curve data and class information from EL image analysis. The idea is, once the machine learning model learns on a large-scale dataset of IV curves and counterpart EL images, it will predict the various defect types of PV modules on the IV curve feature vectors itself with high accuracy without any class information.

Our goal is to predict module-level defects from IV curve features derived with algorithmic and machine learning techniques. Prediction of defect types at module level without an imaging technique such as EL would enable high-speed characterization of modules defects in outdoor field conditions. We provide an overview of the experimental study, data collection, algorithms and computational methods like image processing, feature extraction, feature engineering, feature selection, and machine learning algorithms in Sections 2 and 3. In Section 4, we present the results and discussion of various clustering and classification algorithms. Section 5 contains conclusions of the research.

2 | DATA PROCESSING

The dataset consist of 6051 IV curve and corresponding EL images of field installed 60/72 cell modules of different manufactures at multiple locations across India from 27 projects with age ranging from 1 to 8 years. The bar chart in Figure 1 shows the distribution of plant age when IV–EL measurements were taken.

We collect IV data through an IV curve tracer device along with solar irradiance and module temperature as per methodology described in Section 3 following IEC 61829.[9,10] We have translated our dataset to standard test conditions (STC) (1000 W/m², 25°C) for obtaining standardized IV data as per translation procedure mentioned in IEC 60891.[11] The translation procedure is based on the simplified one-diode model of PV devices. The translation of IV curve as per IEC 60891 correction procedure:

$$I_2 = I_1 \cdot (1 + \alpha_{\mathrm{rel}} \cdot (T_2 - T_1)) \cdot \frac{G_2}{G_1},$$

$$V_2 = V_1 + V_{OC1} \cdot \left\{ \beta_{\mathrm{rel}} \cdot (T_2 - T_1) + a \cdot \ln\left(\frac{G_2}{G_1}\right) \right\} - R'_S \cdot (I_2 - I_1) - \kappa' \cdot I_2 \cdot (T_2 - T_1),$$

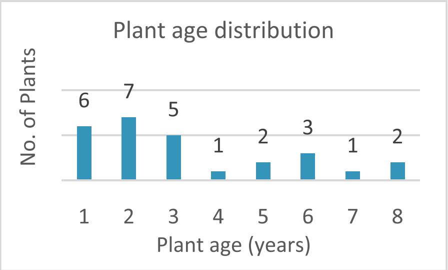


FIGURE 1 Plant age distribution [Colour figure can be viewed at wileyonlinelibrary.com]

where I_1 and V_1 are coordinates of points on the measured IV characteristic; I_2 and V_2 are coordinates of the corresponding points on the corrected IV curve; G_1 and T_1 is the irradiance and temperature as measured with the reference device; G_2 and T_2 is the target irradiance and temperature for the corrected IV characteristic; V_{OC1} is the open-circuit voltage at test conditions; α_{rel} and β_{rel} are the relative current and voltage temperature coefficients of the test specimen measured at 1000 W/m^2; a is the irradiance correction factor for open-circuit voltage which is linked with the diode thermal voltage of the p–n junction and the number of cells n_S serially connected in the module; R'_S is the internal series resistance of the test specimen; and κ' is interpreted as temperature coefficient of the internal series resistance R'_S.

We have extracted the following features from IV curve: maximum power (P_{max}), fill factor (FF), current at maximum power (I_{MP}), voltage at maximum power (V_{MP}), current at short circuit (I_{SC}), voltage at open circuit (V_{OC}), radius of curvature (ROC), near maximum power point (MPP), slope at short circuit (R_{SH}), and slope at open circuit (R_S). We have built linear regression model to estimate the values of R_{SH} and R_S. We have added another feature that is area under the IV curve (AUC_{IV}) which has been calculated with trapezoidal rule. Apart from the mentioned features, few statistical features like skewness, kurtosis, and entropy are also computed based on the IV data distribution. Table 1 and Figure 2 show the different features.

Time difference between measurement of IV curve and EL imaging is only 6–36 h for all the projects. This time duration is very small (~0.01% of module lifetime), which is unlikely to cause any significant pressure impact to reflect changes in PV modules.

Identification of defect types has been done through analysis of EL images. The EL images were analyzed as per the IEC 60904-13 standard.[12] The process by which we capture EL images adds the module orientation variability in the image. We have created the EL image processing pipeline with open-source Python library Open CV. We have used filtering and thresholding technique to reduce noise, perform the perspective correction and rotate the EL images, if required. Annotation of various defect types (as mentioned in Tables 2 and 3) has been performed by two experts from PV Diagnostics. The Cohen's kappa coefficient (κ) has been used in pipeline to measure the interrater reliability score of the annotator. The datasets with substantial good score (>0.85) have been chosen for further steps.

$$\kappa = \frac{p0 - pe}{1 - pe},$$

where $p0$ is relative observed agreement among raters and pe is hypothetical probability of chance agreement.

After annotation, we have classified the EL images in to following classes: cracks (C), solder failures (SFs), PID, and corrosion (PNC). As shunted diode (diode faults) can be easily identified through V_{OC} of module, it is not included in classification. Since PV modules can develop multiple defects, so we have created a total of eight possible combinations shown in Table 4.

After performing the feature engineering step and annotations for all the defect types, we have performed the data preprocessing

TABLE 1 Mathematical formulations for various features of IV data

Feature	Definition
I_{SC}	Current at short circuit
V_{OC}	Voltage at open circuit
P_{max}	$P_{max} = \max_{j \in [N]} (V_j \cdot I_j)$, where $N = \{1, 2, \ldots, n\}$ and n is the total number of data points in IV curve
FF	$FF = \frac{P_{max}}{V_{OC} \cdot I_{SC}}$
V_{MP}, I_{MP}	$V_{MP}, I_{MP} = V[idx_{max}], I[idx_{max}]$ $idx_{max} = \text{index of} \max_{j \in [N]} (V_j \cdot I_j)$
ROC	$ROC = \left\lvert \frac{(\dot{x}^2 + \dot{y}^2)^{\frac{3}{2}}}{\dot{x}\ddot{y} - \ddot{x}\dot{y}} \right\rvert$, where $x = V$, $y = I$, and calculation is done on few IV data points around MPP
R_{SH}	$\frac{1}{R_S} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, where $x = V$, $y = I$, and calculation is done on few IV data points around V_{OC}
R_S	$\frac{1}{R_{SH}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, where $x = V$, $y = I$ and calculation is done on few IV data points around I_{SC}
AUC_{IV}	$AUC_{IV} = \sum_{j=1}^{n-1} (V_{j+1} - V_j) \cdot \frac{I_{j+1} + I_j}{2}$
$I_{skewness}$	$I_{skewness} = \frac{\sqrt{(N-1) \cdot N}}{N-2} \cdot \frac{m_3}{m_2^{1.5}}, m_i = \frac{1}{N} \cdot \sum_{j=1}^{N} (y_j - \bar{y})^i$
$V_{skewness}$	$V_{skewness} = \frac{\sqrt{(N-1) \cdot N}}{N-2} \cdot \frac{m_3}{m_2^{1.5}}, m_i = \frac{1}{N} \cdot \sum_{j=1}^{N} (x_j - \bar{x})^i$
$I_{kurtosis}$	$I_{kurtosis} = \frac{m_4}{\sigma^4}$
$V_{kurtosis}$	$V_{kurtosis} = \frac{m_4}{\sigma^4}$
$I_{entropy}$	$I_{entropy} = -\sum_{j=1}^{N} (P(I_j) \cdot \ln P(I_j))$
$V_{entropy}$	$V_{entropy} = -\sum_{i=1}^{N} (P(V_i) \cdot \ln P(V_i))$

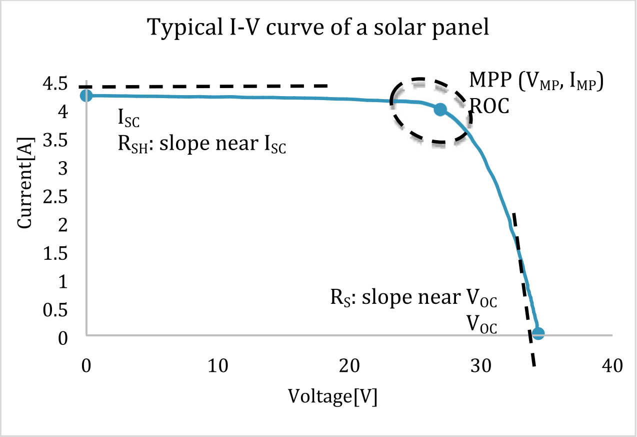


FIGURE 2 Typical IV curve of a PV module showing major parameters [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Examples of defects in EL images[13]

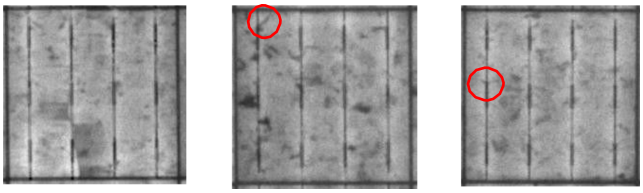
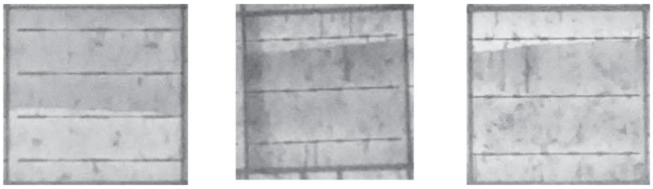
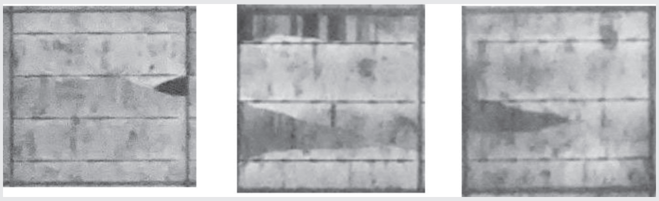
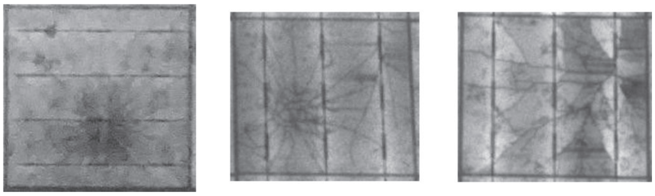
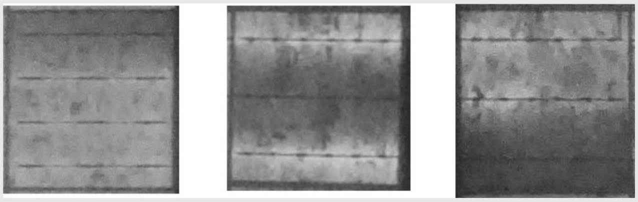
Defect	Comment	Image
Microcrack	A fracture that does not extend more than 2 cm from origin.	
Partial crack	A single fracture that does not extend the full distance across the cell.	
Single or full crack	A single fracture that extends the full distance across the cell. It is only one fracture. It does not have any sharp turns.	
Compound crack	A fracture that has more than one crack or has crack with sharp turns. With cycling, a sharp point will likely turn into another crack. May result from localized pressure.	
Point/spider crack	A point fracture has a center of originating position. Several fractures radiate out from the center. Point fractures may also be referred to as spider fractures due to their web like appearance.	
Solder failure	Indicated by areas of low luminescence surrounding a solder joint. Solder joint failures increase the resistance of the cell, thus reducing performance and may result in hot spots.	

TABLE 3 Defect annotations on EL images

Sr. No.	Defect type
1	Cracks (micro, partial, full, compound/spider)
2	Solder failure (partial, full)
3	Potential-induced degradation (PID)
4	Back sheet scratch, corrosion
5	Diode fault

TABLE 4 Defect class label information of EL images

Defect class	Class definition
ALL	All defects class (C, SF, and PNC)
False	No defects
C	Cracks
SF	Solder failure
PNC	PID and corrosion
C_SF	Cracks and solder failure
C_PNC	Cracks and PID
SF_PNC	Solder failure and PID

where we have dealt with missing values, outliers, and data normalization and standardization using the open-source python pandas library to prepare the dataset for classification. This process is shown in Figure 3.

3 | MODELING AND ANALYSIS

All the feature vectors of IV data and class information of EL data have been combined to form a new dataset for machine learning analysis as show in final step of Figure 3. The preprocessing steps have been performed including data cleaning, normalization, and standardization. Further, class imbalances have been identified. The over sampling method with Synthetic Minority Oversampling Technique (SMOTE) has been implemented to deal with class imbalance problem. *t*-Distributed stochastic neighbor embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP) unsupervised machine learning have been implemented for visualization and to find out if the classes are well separated in lower embedded space for selected features. If classes are well separated, then the results were added to feature list to improve the classification models. Feature importance and selection approach have been implemented to improve the model's accuracy and reduce the complexity with random forests (RF) and Extreme Gradient Boosting (XGBoost) to find out the top features. The RF and XGBoost have been implemented with hyperparameter of large parameter space and selected the top model with respect to accuracy performance metric (refer Figure 4 for complete process flow and implementation steps).

We have first performed the analysis on dataset with unsupervised machine learning algorithms for dimension reduction. The *t*-SNE and UMAP have been utilized for this research.[14,15] We implemented the dimension reduction of all feature vectors in two- and three-dimensional maps for visualization. This analysis was performed by tuning the hyperparameters of the *t*-SNE and UMAP on the clusters found in lower embedded space against the labels of defect types. The code was written using Python with its additional packages of NumPy, Pandas, scikit-learn, umap-learn, SciPy, and matplotlib.[16,17]

3.1 | Unsupervised machine learning

t-SNE is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. The technique can be implemented via Barnes–Hut approximations, allowing it to be applied on large real-world datasets. *t*-SNE minimizes

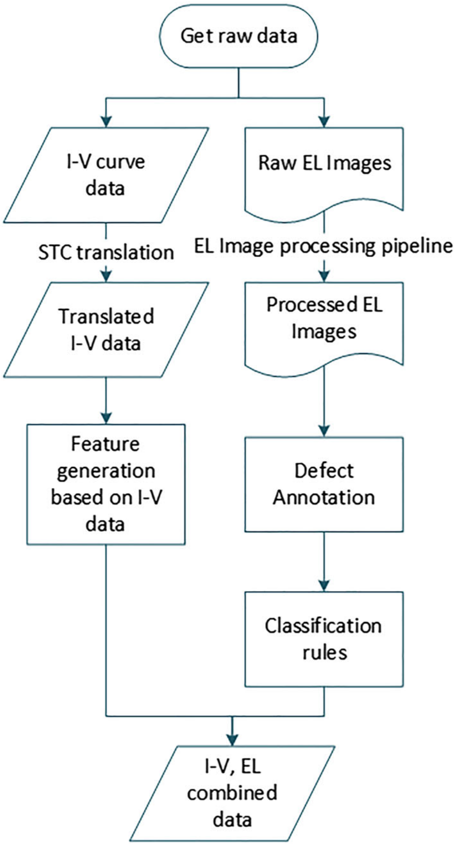


FIGURE 3 Process flow for IV and EL data including data preprocessing steps, STC translation feature generation, image processing pipeline, defect annotation, and classification rules [Colour figure can be viewed at wileyonlinelibrary.com]

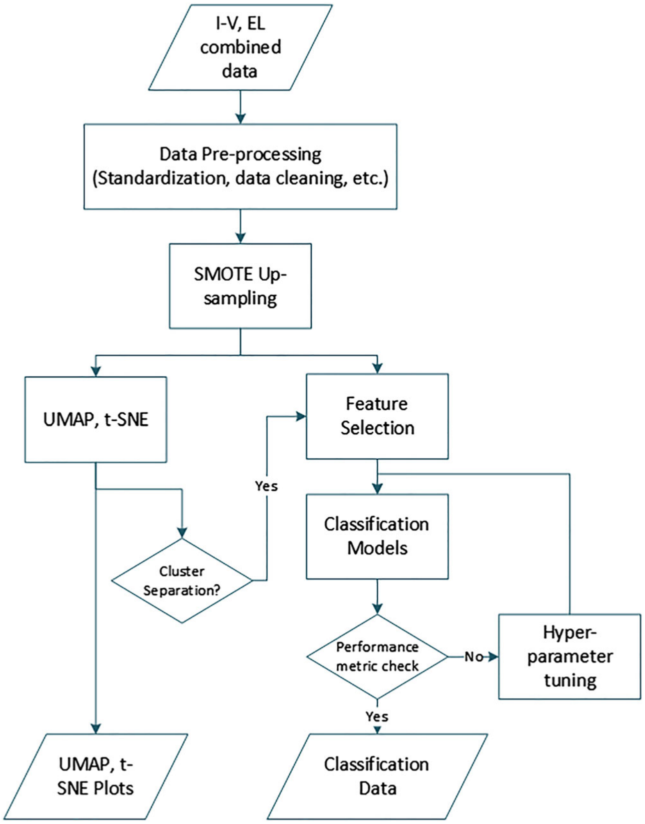


FIGURE 4 Flow diagram showing various steps of data modeling and machine learning development [Colour figure can be viewed at wileyonlinelibrary.com]

the divergence between two distributions (a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding).[14]

UMAP is a novel manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that is applicable to real-world data. The UMAP algorithm is competitive with *t*-SNE for visualization quality and preserves more of the global structure with superior run time performance. Furthermore, UMAP has no computational restrictions on embedding dimension, making it viable as a general-purpose dimension reduction technique for machine learning.[15]

We have performed the analysis on 15 features vectors with unsupervised machine learning for dimension reduction of sample size of 6051 data points with *t*-SNE. This analysis was performed by tuning of hyperparameters of *t*-SNE on the clusters found in lower embedded space against the labels of defect types. We used different hyperparameters to tune the model such as distance metric: "Canberra," "Dice," "Manhattan," "Chebyshev," perplexity ranging from 15 to 50 with a step size of 5, learning rate fixed at 200, and number of iterations were 1000. The results of *t*-SNE algorithm indicate that there is no significant clusters separation. But when we separate only the defects into clusters with Canberra distance metric, minor cluster formation can be seen.

We also used UMAP unsupervised machine learning for dimension reduction for same data points. The analysis was performed by tuning of hyperparameters of UMAP on the clusters found in lower embedded space against the labels of defect types. We used different hyperparameters to tune the model such as distance metric: "Euclidean," "Manhattan," "Chebyshev," "Hamming," "Dice," local neighborhood from 10 to 25, learning rate fixed at 1, and number of iterations were 500. The results of UMAP algorithm indicate that there is no significant clusters separation.

3.2 | Machine learning classifiers

RF is an ensemble of ML techniques that builds multiple decision tree classifiers on random subsamples of the training dataset. Each decision tree predicts the response by following the tree's decisions from the root to the leaf. The output of each decision tree is then averaged to determine the prediction.[18] RF's main advantage is leveraging the power of many randomly selected trees to represent the solution. Thus, instead of using one decision tree, RF uses all the decision trees to determine the classification; this procedure reduces errors and uncertainties.[19] In this study, the number of trees selected was ranging from 25 to 1000 in step size of 25. The minimum number of samples that are required to split an internal node is set to 25. The maximum depth of the tree is set from 10 to 200.[20] The training algorithm for RF applies the general technique of bootstrap aggregating or bagging to tree learners.

The other classification ML algorithm we have used is XGBoost. XGBoost was developed by Chen and Guestrin,[21] a scalable tree boosting system that is widely used by data scientists, providing state-of-the-art results on many problems. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the gradient boosting framework. It provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. We have chosen XGBoost classifier because it can also help us deal with large-scale datasets and easily run on distributed environment (Hadoop, SGE, and MPI).

Open-source Python was utilized in XGBoost implementation for classification of defect types with feature vectors. Various metrics were considered to determine the goodness of fit of these models. These metrics include accuracy, F1 score, precision-recall curve, number of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) at various decision thresholds and calibration curves of the estimated probabilities, as shown in Table 5.

The recall metric measures the percentage of total relevant results correctly classified by the algorithm, while precision is the ratio between correctly labeled positive outcomes and the total predicted positive outcomes. Accuracy is defined as the strength of the correlation between the predicted and the actual labels.[19] Accuracy is the ratio between the number of correct predictions and the total number of predictions. Nevertheless, accuracy is not the best representation of performance on unbalanced datasets. Hence, the F1 score metric, defined as the harmonic mean of precision and recall, is also computed in this study. It has been shown that the F1 score is a better indicator when analyzing unbalanced datasets.[22]

3.3 | Feature selection and hyperparameter tuning

All the initially computed features as mentioned in Table 1 are selected except R_{SH} for initial classification models. The dataset is divided into training and validation in a ratio of 80:20. The initial model has been built using this split of data, and we have used five-fold cross validation to confirm the stability of algorithm. The initial training accuracy was ∼80% while validation accuracy of ∼75%. Based on the confusion matrix of initial results, we dropped class C, class C_SF, and class ALL as the accuracy values for these defects were significantly lower than other classes.

Afterwards, we built the RF and XGBoost classification models using only selected classes and computed the feature importance for

TABLE 5 Classification performance metrics and their definition

Performance metric	Accuracy (%)	Recall, R (%)	Precision, P (%)	F1 score (%)
Formulae	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$	$\frac{2PR}{P+R}$

both models. Then, we have arranged the features in decreasing order of importance as show in Figures 5 and 6. Next, we have selected the top five features to further enhance the model performance. To study the influence of hyperparameters on the RF classifiers' performances, combinations of hyperparameters are used, with each combination leading to one classifier (refer Table 6). Similarly, for XGBoost grid search was performed by selecting the hyperparameters sets mentioned in Table 6 below to further optimize the model.

4 | RESULTS AND DISCUSSION

The feature set V1 comprises all feature vectors except the R_{SH} feature. While feature vector set V2 comprises of top five selected feature vectors as per feature importance scores calculated for both classification models, respectively. We have found that the V2 set of RF model includes R_S, ROC, FF, $V_{kurtosis}$, and V_{OC} feature vectors as shown in Figure 5. Similarly, for XGBoost model, V2 set includes $V_{skewness}$, $V_{kurtosis}$, FF, V_m, and AUC_{IV} feature vectors as shown in Figure 6. We have compared the performance scores on validation set using both feature vector sets V1 and V2, when they are used as inputs for both ML classifiers RF and XGBoost as shown in Table 7. For both cases, vector (V2) outperforms vector (V1) in all performance measures, achieving higher accuracy and F1 score. A significant difference can be observed between RF and XGBoost classifiers. It has also been observed through a comparison between the training and validation sets that the data has not been overfitted.

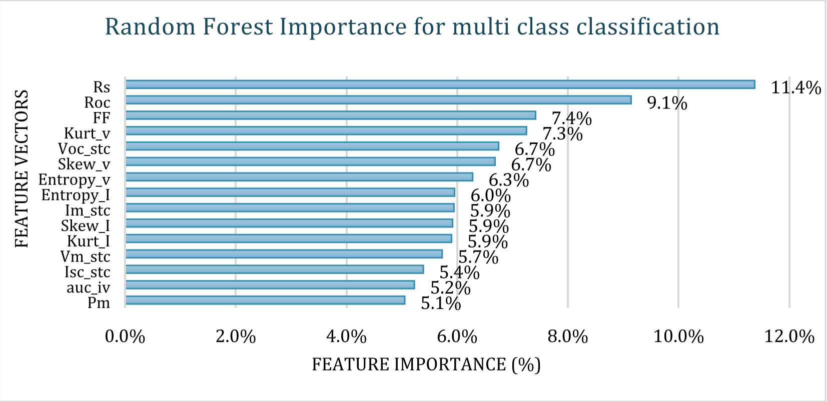


FIGURE 5 Feature vector importance ranked from highest to lowest for RF classifiers [Colour figure can be viewed at wileyonlinelibrary.com]

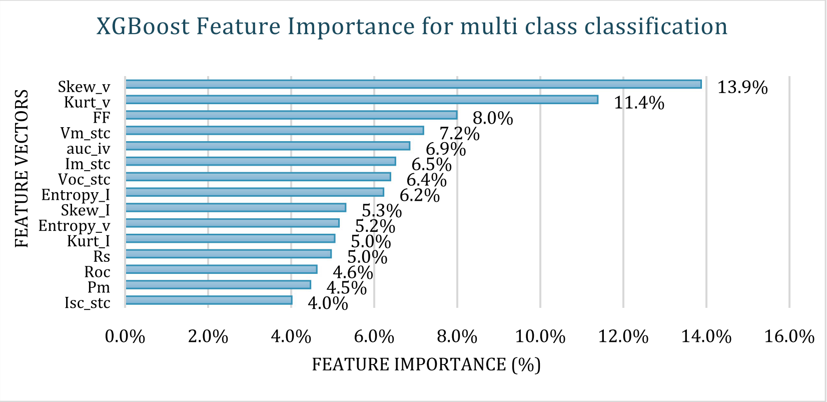


FIGURE 6 Feature vector importance ranked from highest to lowest for XGBoost classifier [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 6 Hyperparameters for RF and XGBoost models

RF parameters	XGBoost parameters
Maximum depth ∈ {10, 20, ..., 200}	Maximum depth ∈ {5, 10, 15}
Number of trees ∈ {25, 50, ..., 1000}	Number of trees ∈ {500, 1000, 1500, 2000}
Split criterion ∈ {Gini, entropy}	Mini split loss ∈ {0, 0.5, 2, 10}
Max feature ∈ {sqrt, log2}	Learning rate ∈ {0.01, 0.1, 0.3, 0.5}

TABLE 7 Overall results of feature vectors sets with RF and XGBoost classifiers

Feature vectors	Feature vectors V1		Feature vectors V2	
ML classifiers	RF	XGBoost	RF	XGBoost
F1 score (%)	88.8	89.0	91.2	97.0
Accuracy (%)	88.6	92.3	90.8	95.6
Recall (%)	92.0	93.7	93.4	98.0
Precision (%)	88.7	89.6	91.3	96.0

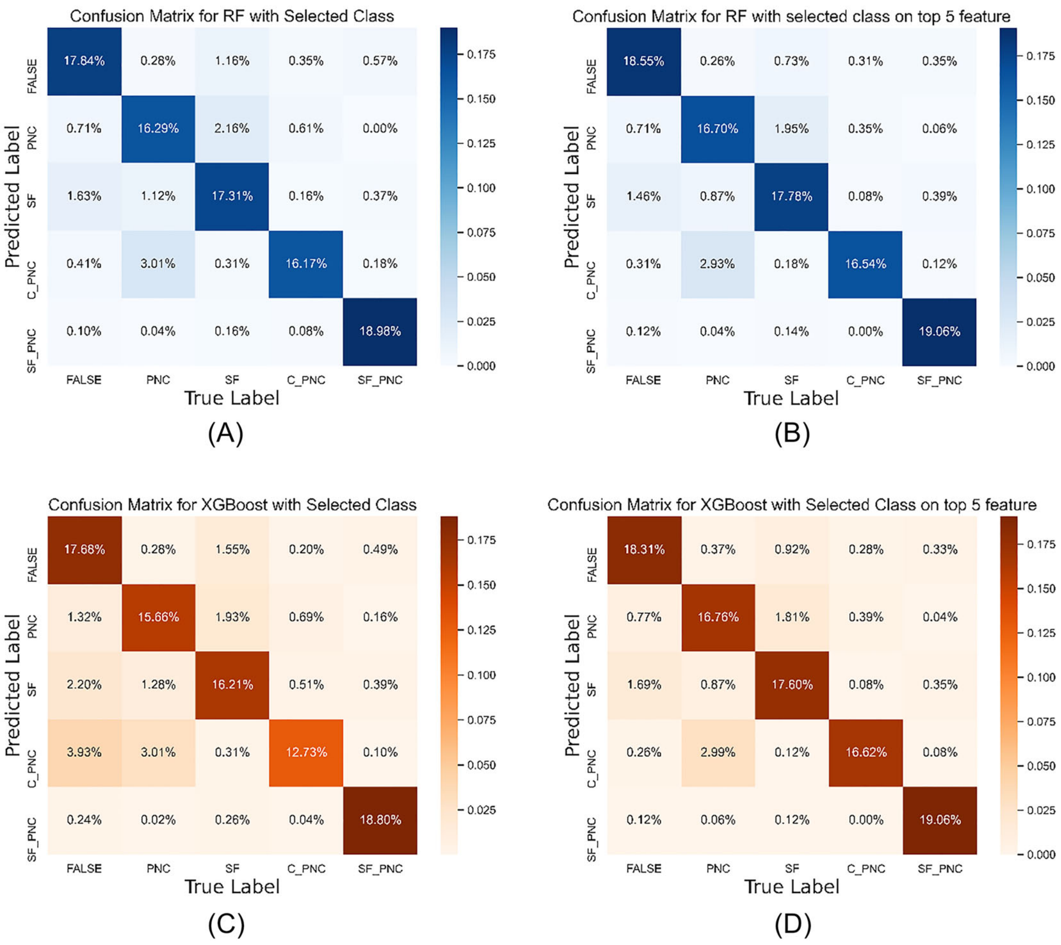


FIGURE 7 Confusion matrix for (A) RF with all feature vectors set V1, (B) RF with top five feature vectors set V2, (C) XGBoost with all feature vectors set V1, and (D) XGBoost with top five feature vectors set V2 [Colour figure can be viewed at wileyonlinelibrary.com]

We note that V2 feature vector and XGBoost classifier is the best combination for performance evaluation, achieving 95.6% accuracy. Figure 7 provides an overall performance evaluation of V1 and V2 and highlights the correlation between the actual and predicted labels. The percentage of PV module defects predicted incorrectly in the different class categories is significantly higher for V1. Even though V1 results in 93.7% recall value, using V2 with the top features improves the performance to an overall recall value of 98% for the XGBoost classifier. As expected, the V2 feature vector achieved a precision value of 96% (XGBoost classifier), outperforming the combined approach (V1) which resulted in 89.6%.

5 | CONCLUSION

The early phase detection of defects as cracks, SF, PID, and finger interruption in solar cells is important to estimate power generation and life of PV modules. Analyzing EL images to locate and identify these failures is typically a time-consuming manual process and requires expert knowledge. In this paper, a machine learning-based defect identification method was presented for IV–EL correlation. The technique uses EL images for the class information to classify three major classes of defects using feature vectors based on IV curve parameters and statistical features. It has been shown that defects in EL images correlate well with trends seen in IV curve data. The developed feature vector achieves an accuracy and F1 score which are quite promising for defects identification in PV Industry. While performing the model training–testing, we observed that the overall accuracy and other performance metrics decreased with inclusion of crack defect class. This is because the impact of cracks on IV data is very complex to analyze, and it also depends on severity levels of different types of cracks. Thus, in future, with increased dataset, the accuracy can be further increased for the dropped classes too.

As the proposed method requires less computation power and time, it will be valuable for outdoor IV inspection including handled IV inspection, Internet of things (IoT)-based cloud inspection through module monitoring. The learned models can be easily integrated on a software as a service (SaaS) platform and deployed on cloud for processing large-scale datasets as XGBoost provides a parallel tree boosting and its codes can run on distributed environment (Hadoop, SGE, and MPI).

ACKNOWLEDGEMENTS

The authors thank PV Diagnostics for providing all the R&D and logistical support. PV Diagnostics has provided a constant support throughout this process. We would also like to express our sincere thanks to the operations team, backend analysis team, and the EL image analysis experts who were instrumental in collecting, preprocessing, annotating, and reviewing the dataset. This helped the authors to build optimal ML models for defect identification using IV–EL data. We would also express our gratitude to Mr. Ankur Gahlot who helped us in the inception of this analysis.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Haegel N, Atwater H, Barnes C, et al. Terawatt-scale photovoltaics: tranform global energy. *Science*. 2019;364:836-838.
2. Sharma V, Chandel S. Performance and degradation analysis for long term reliability of solar photovoltaic systems: a review. *Renew Sustain Energy Rev*. 2013;27:753-767.
3. Eitner U, Kajari-Schrder S, Kntges M, Altenbach H. Thermal stress and strain of solar cells in photovoltaic modules. In: *Shell-Like Structures: Non-Classical Theories and Applications*. Springer Berlin Heidelberg; 2011:453-468.
4. Kongtes M, Kurtz S, Jahn U, et al. *IEA-PVPS T13-01 Performance and Reliability of Photovoltaic Systems: Review of Failures of PV Modules*. International Energy Agency; 2014.
5. Jordan D, Silverman J, Wohlgemuth J, Kurtz S, VanSant K. Photovoltaic failure and degradation modes. *Prog Photovolt: Res Appl*. 2017;25(4):318-326.
6. Pletzer T, Molken JV, Ribi S, Breitenstein O, Knoch J. Influence of cracks on the local current–voltage parameters of silicon solar cells. *Prog Photovolt*. 2015;23(4):428-436.
7. Trupke T, Nyhus J, Haunschild J. Luminescence imaging for inline characterisation in silicon photovoltaics. *Phys Status Solidi (RRL)*. 2011;5(4):131-137.
8. Fada J. Correlation of I–V curve parameters with module-level electroluminescent image data over 3000 hours damp-heat exposure. IEEE Photovolt Spec Conf. 2017;2697-2701.
9. International Electrotechnical Commission. IEC 60904-1 Photovoltaic devices—Part 1: measurement of photovoltaic current-voltage characteristics, 2020.
10. International Electrotechnical Commission. IEC 61829 Photovoltaic (PV) array—on-site measurement of current-voltage characteristics, 2015.
11. International Electrotechnical Commission. IEC 60891 Photovoltaic devices—procedures for temperature and irradiance corrections to measured I-V characteristics, 2009.
12. International Electrotechnical Commission. IEC TS 60904-13 Photovoltaic devices—Part 13: electroluminescence of photovoltaic modules, 2018.
13. Daniels E. Suncycle USA. 2018. [Online]. Available: https://www.suncycleusa.com/.
14. Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579-2605.
15. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction; 2020.
16. Linge S, Lagtangen H. *Programming for Computations-Python: A Gentle Introduction to Numerical Simulations with Python*. Springer; 2016.
17. McKinney W. pandas: a foundational Python library for data analysis and statistics. *Python High Perform Sci Comput*. 2011;14(9):1-9.
18. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
19. Ziegler A, James R, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. Springer; 2016.
20. Breiman L, Freidman J, Stone C, Olshen R. *Classification and Regression Trees*. CRC Press; 1984.
21. Chen T, Guestrin C. XGBoost: a scalable tree boosting system, In 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016;785–794.
22. Powers D. Evaluation from precision, recall and f-factor to ROC, informedness, markedness, and correlation. *Mach Learn Technol*. 2008;2:37-63.

How to cite this article: Kumar V, Maheshwari P. Advanced analytics on IV curves and electroluminescence images of photovoltaic modules using machine learning algorithms. *Prog Photovolt Res Appl*. 2021;1-9. doi:10.1002/pip.3469