

ARTICLE OPEN



Extracting bulk defect parameters in silicon wafers using machine learning models

Yoann Buratti no¹ , Quoc Thong Le Gia¹, Josef Dick¹, Yan Zhu¹ and Ziv Hameiri¹

The performance of high-efficiency silicon solar cells is limited by the presence of bulk defects. Identification of these defects has the potential to improve cell performance and reliability. The impact of bulk defects on minority carrier lifetime is commonly measured using temperature- and injection-dependent lifetime spectroscopy and the defect parameters, such as its energy level and capture cross-section ratio, are usually extracted by fitting the Shockley-Read-Hall equation. We propose an alternative extraction approach by using machine learning trained on more than a million simulated lifetime curves, achieving coefficient of determinations between the true and predicted values of the defect parameters above 99%. In particular, random forest regressors, show that defect energy levels can be predicted with a high precision of ±0.02 eV, 87% of the time. The traditional approach of fitting to the Shockley-Read-Hall equation usually yields two sets of defect parameters, one in each half bandgap. The machine learning model is trained to predict the half bandgap location of the energy level, and successfully overcome the traditional approach's limitation. The proposed approach is validated using experimental measurements, where the machine learning predicts defect energy level and capture cross-section ratio within the uncertainty range of the traditional fitting method. The successful application of machine learning in the context of bulk defect parameter extraction paves the way to more complex data-driven physical models which have the potential to overcome the limitation of traditional approaches and can be applied to other materials such as perovskite and thin film.

npj Computational Materials (2020)6:142; https://doi.org/10.1038/s41524-020-00410-7

INTRODUCTION

In the wake of the climate crisis, transitioning to a carbon-free society is crucial in creating a sustainable future. In their 2018 report, the Intergovernmental Panel on Climate Change (IPCC) explores sustainable scenarios with PV systems producing up to a third of the planet's renewable energy¹. Today's most prominent PV technology is silicon (Si) based solar cells². Improving the efficiency and reliability of Si PV modules and solar cells, as well as reducing the cost of this technology, is paramount in meeting the IPCC targets³. Research focused toward higher efficiency Si solar cells has shone light on Si bulk defects as a key contributor to their total efficiency loss⁴. It is therefore primordial to identify and characterize bulk defects in order to eliminate or at least reduce their impact. Techniques such as temperature- and injectiondependent lifetimes spectroscopy (TIDLS)⁵ enables characterization of a wide range of Si bulk defects, such as copper⁶, aluminium^{7,8}, chromium^{9,10}, iron^{11,12} and many more^{13–17}. To identify defects, the Shockley-Read-Hall (SRH) equation 18,19 that describes the impact of a defect on the overall minority carrier lifetime is often used to fit the lifetime measurements²⁰ defect parameters are defined by the SRH equation to represent a defect: its energy level Et and its electron and hole capture crosssections, σ_n and σ_p , respectively. The state of the art fitting method uses Rein's defect parameter solution surface (DPSS)²⁰ which determines, for each injection-dependent measurement at a fixed temperature, a set of solutions to the SRH equation²⁰. By continuously varying E_{t} , the best fit for other parameters, or combinations of parameters, can be found, creating a continuous surface in the parameter space which represents the entirety of the mathematical solutions for that fit. By repeating the process for each temperature-dependent measurement, different DPSS curves are determined. These curves usually intersect at two points that indicate the possible combinations of the defect parameters, with one set of parameters in the upper bandgap and one in the lower bandgap^{7–13,21}. This method, henceforth referred to as the 'traditional approach', has been refined over the years, such as linearizing the SRH equation under specific conditions^{14,22} or applying faster convergence techniques to find the intersection points²¹. However, the technique and its derivatives assume that the defect parameters are independent of temperature and that the defect possesses only one energy level in the Si's bandgap^{7,9}. In addition, in practice, it fails to identify which of the two mathematical solutions is the correct one.

To overcome these limitations, this study proposes a machine learning (ML) based approach to extract the defect parameters from lifetime curves. ML-based methods are already used at the PV system level, for example for fault detection 23,24 or to identify cracks in modules using luminescence imaging techniques^{25–28}. ML has also been used in non-Si applications to find relevant material parameters for fabrication of CIGS solar cells²⁹, multijunction solar cells³⁰, organic solar cells³¹, or perovskite solar cells³². More recently, Kurchin et al.³³ proposed a Bayesian-based model to predict the probability distribution of the defect parameters from temperature dependent current-voltage measurements. The model is designed to fit a single combination of defect parameters and needs to be retrained for different measurement scenario, such as iron or copper in silicon. Furthermore, it seems that the use of current-voltage measurements to extract the defect parameters can be influenced by other solar cells parameters, such as edge recombination and shunt resistance, that can mask the impact of the studied defect. This study's approach brings the technological prowess of ML to defect parameters extraction in order to propose an alternative way to



¹The University of New South Wales, Sydney, NSW 2052, Australia. [⊠]email: yoann.buratti@student.unsw.edu.au

solve this decades-old challenge. Multiple ML training algorithms are considered, and compared, to predict the defect parameters directly from TIDLS measurements with no knowledge of the SRH equation. The developed method is not limited to silicon and can be used for non-Si PV materials, such as perovskites or CIGS.

RESULTS

Defect parameter regression

Five iterations of training for each of the ML training algorithms are performed. Each training uses different random model initialization and random training/validation dataset split to ensure the repeatability of the ML algorithm's predictive power. For neural networks (NN), the evaluation data is held out until the end of the epoch-based training. In Fig. 1, the results of each training's coefficient of determination (R^2) score are presented for the different ML models for both $E_{\rm t}$ (a) and k (b) predictions. The models have been ordered from the best performers to the worst and the average and standard deviation R^2 are given for each model and regression target. In the figure, the statistical box plot for each model is shown in green, while the black dots represent each individual training iteration.

Apart from the support vector machine (SVM) models – expected to perform better when the datasets are small and sparse³⁴—and one instance of NN, all trained ML models have am R^2 above 95% on the validation dataset, showing the feasibility of using ML as an inverse function, with no a priori knowledge of the SRH equation. The highest average R^2 (>99%) is achieved by the

random forest (RF) model for both k and $E_{\rm t}$ prediction. From Fig. 1a, b, it can be concluded that the decision tree methods, such as RF, adaptive boosting, and gradient boosting achieve the highest average R^2 score in predicting both $E_{\rm t}$ and k, followed by NN and SVM. NN have a higher variance than the other models, due to its sensitivity to the random initial weights and the random order in which the data is processed during training³⁵. Indeed, the training of NN relies on backpropagation steps that shift the weights of the network proportional to the error of the data points that it is trained to predict in order to minimize that error³⁶. Changing the order in which the NN evaluates the data shifts the networks to different local minima because of these initial larger weight shifts.

Figure 2 shows the predicted parameters versus the true parameters for the best performing RF model using the validation dataset. Both (a) E_t and (b) k predictions have R^2 scores above 99%. In each graph, the data points are semi-transparent, and the darker area outlines the linearity of the correlation between predicted and true values, as most data points lie on the y = x line. The RF model correctly predicts E_t across the entire range 87% of the time at ± 0.02 eV precision which is typical error bars for reported E_t values (between ± 0.02 eV to ± 0.04 eV)^{8–10,15–17}. It is noticeable that far from the mid-gap ($|E_t| > 0.1$ eV), the correct predication is much higher (91%) for this precision level. Close to the mid-gap ($|E_t|$ < 0.1 eV), the prediction of E_t at ±0.02 eV correctness drops to 75%. This can be explained by examining the SRH equation [Eq. (2)] when E_t is close to 0 eV. The dependency on E_t comes from the n_1 and p_1 terms [Eq. (4)]. When E_t approaches 0 eV, both n_1 and p_1 terms becomes negligible compare to $n_0 + \Delta n$

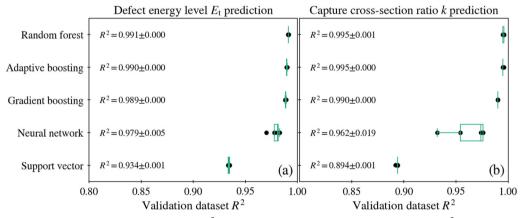


Fig. 1 Defect parameter regression model comparison. a R^2 scores of models on validation dataset for E_t . b R^2 scores of models on validation dataset for k.

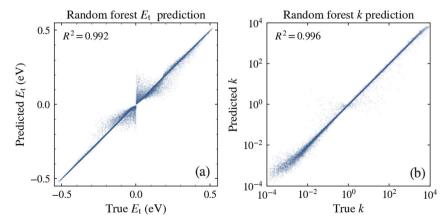


Fig. 2 Random forest defect parameter regression. a True vs predicted E_t on validation dataset for best performing random forest. b True vs predicted k on validation dataset for best performing random forest.

and $p_0 + \Delta n$, respectively. Thus, the resulting SRH lifetime has a weak dependence on E_t , increasing the difficulty of both human and ML prediction near the mid-gap.

To validate this explanation, a simulation of 200 defects with a fixed k=100 and $E_{\rm t}$ spanning from -0.1 eV to 0.1 eV is performed. Figure 3 presents the lifetime curves of these defects at 400 K. It can be observed that the defects are indistinguishable from one another on a macro scale. The insert is a zoom-in of the ensemble of lifetime curves at low injection level and it reveals that this ensemble of curves is 50 ns wide. Whereas it is extremely difficult to distinguish those lifetime curves with the naked eye, the RF model still correctly predicted 75% of the defects with ± 0.02 eV accuracy.

Defect bandgap location classification

In the traditional DPSS approach two possible solutions are often identified, one in the lower bandgap and one in the upper bandgap^{7–9,21}. In certain cases, the half bandgap location can be determined using additional measurements, such as deep level transient spectroscopy¹⁷. To address this limitation, ML models are trained to classify in which half bandgap the defect is located. The models have been ordered from the best performers to the worst one in Fig. 4a, and the average and standard deviation of each model's accuracy. The statistical box plot for each model is shown in green, while the black dots represent each individual training iteration. NN significantly outperforms the other models, reaching an average accuracy of 82.7%. It is worth noting that all the other

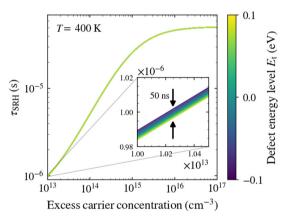


Fig. 3 Mid-gap defects lifetime curve simulation. Simulation of 200 defects lifetime curves for $E_{\rm t}$ near the mid-gap and k=100 at 400 K. The insert is a zoom-in to estimate the width of the lifetime 'band'.

algorithms produce models above 75% accuracy proving the possibility of identifying the half bandgap's location with TIDLS measurements. Despite NN's good performance, the variation in the accuracy score between trained models shows the same behaviour than in the regression's case of sensitivity to the first few steps of backpropagation (as shown by the standard deviation of this model). From Fig. 4a, it can also be observed that the RF models are significantly better than its boosting counterparts. Boosting techniques add error scoring methods during the training of the different decision trees of the RF³⁷. The difference in performance between boosting and RF is due to the overfitting of the training data by the boosting models³⁸. The best trained NN has an accuracy of 87.7% on the validation dataset and its confusion matrix is shown in Fig. 4b. The confusion matrix is symmetric, as expected, showing that both half bandgaps are treated equally.

To provide more insight into the 12.4% incorrectly predicted half bandgap locations, the corresponding defects are plotted in Fig. 5, as a k vs $E_{\rm t}$ plot. Each data point, representing a single defect parameter combination fed into the NN model, is coloured based on the NN's classification probability towards their true half bandgap location. The probabilities are obtained directly from the output layer of the NN, which are known to be well-calibrated 39 . The grey datapoints corresponds to correctly predicted defect half bandgap location while coloured datapoints to incorrectly predicted defect half bandgap location. The colour scale represents the correct label probability, and for incorrectly labelled datapoints, is necessarily between 0 and 0.5. The darker the datapoints, the more inaccurate the classification by the NN is.

The data points aggregate around two trends. For defects with $|E_t| < 0.1$ eV, the classification accuracy drops to 68.7%, which still provides a decent indication regarding the half bandgap location without the need for additional measurements. As previously discussed for Fig. 2, defects with E_t near the mid-gap are harder to differentiate for both human and ML. Outside the mid-gap range ($|E_t| > 0.1$ eV), the NN classification model performs extremely well, with a 94.7% half bandgap location classification accuracy. The second trend is that outside the mid-gap region, the most incorrect classification is for defects with k values around 0.83 \pm 0.01, represented by the dotted black line in Fig. 5. This corresponds to the ratio between hole and electron thermal velocity. Indeed, when the value k approaches v_p/v_n , it equalizes τ_{n0} and τ_{p0} , resulting in the following simplified SRH equation:

$$\tau_{SRH} = \tau_{n0} \cdot \left(1 + \frac{p_1 + n_1 + \Delta n}{p_0 + n_0 + \Delta n} \right) \tag{1}$$

Since n_1 and p_1 have opposite E_t dependence, the SRH equation in the case $k = v_p/v_n$ is symmetric in respect to E_t . This means that for a fixed k value, satisfying $k = v_p/v_n$, or equivalently $\sigma_p/\sigma_p = v_p/v_n$,

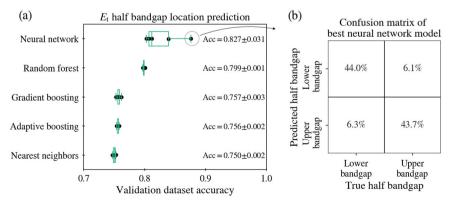


Fig. 4 Defect half bandgap energy location classification model comparison. a Accuracy scores for each model on validation dataset. b True vs predicted confusion matrix on validation dataset for best performing neural network.

npj

the lifetime curves for any given $E_{\rm t}$ are identical to the lifetime curves of the opposite energy level $-E_{\rm t}$. For this reason, it is difficult to pinpoint the exact half bandgap location of defects when the value of $k:=\sigma_{\rm D}/\sigma_{\rm D}$ is close to $v_{\rm D}/v_{\rm D}$, as shown in Fig. 5 where

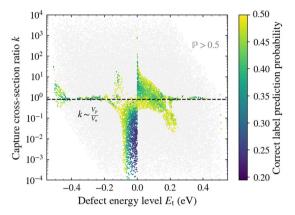


Fig. 5 Best performing neural network k vs $E_{\rm t}$ plot. Coloured datapoints represents defects where half bandgap location was incorrectly predicted. Grey datapoints represents correctly predicted half bandgap location. The dotted black line corresponds to k values close to $v_{\rm p}/v_{\rm n}$.

the prediction probability tends to 50% for each half. However, the regression targets k and E_t are not compromised by that regime, as seen in Fig. 2.

Verification

A comparison between DPSS and the best-trained ML models is given in Fig. 6. The SRH lifetime curves at various temperatures for a selected defect ($E_{\rm t}=0.335$ eV and k=1.1) are presented in Fig. 6a. From these, five DPSS curves are generated following Rein's method²⁰ [see Fig. 6b] to identify the defect parameters. Figure 6b also includes the combined predications of $E_{\rm t}$ and k by the best of the previously trained ML regressor models of each category (indicated by coloured symbols). As can be noticed, a great agreement is demonstrated between all the ML models and the DPSS method in terms of the prediction of the defect parameters. The best regressor predict $E_{\rm t}$ within ± 0.0003 eV and k within ± 0.05 of the simulated values.

As stated previously, the DPSS approach usually provides two solutions for the defect parameter extraction, with one in the upper half bandgap and one in the lower half bandgap. To distinguish between the two solutions, the sharpness of the intersection of the DPSS is plotted in Fig. 6c as a function of E_t , where the sharpness ε is defined as the standard deviation of the k-DPSS curves divided by the mean of the curves. The sharpness is only defined when all the DPSS curves exists, thus, the discontinuity in the sharpness curve. Using this approach, the

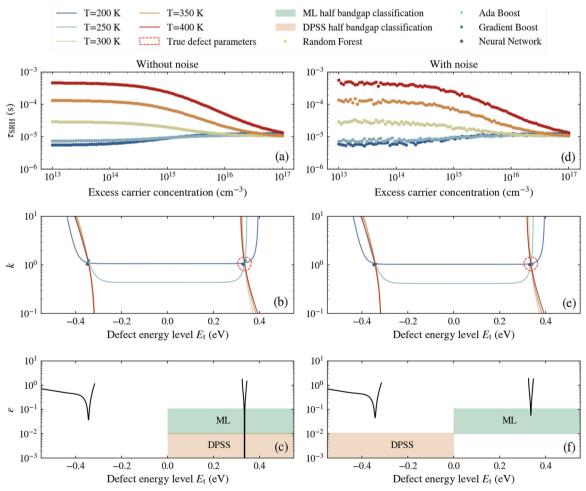


Fig. 6 Traditional and proposed method comparisons on simulated defect. Comparison of the same defect parameters ($E_t = 0.335$ eV and k = 1.1) without (a–c) and with (d–f) noise. a, d Simulated SRH lifetime curves. b, c Calculated DPSS curves for $k := \sigma_n/\sigma_p$. c, f DPSS curves intersection sharpness ε defined as the standard deviation of the k-DPSS curves divided by the mean of those curves.

minimum of the sharpness curve points to the most probable solution. In Fig. 6e, the DPSS curves sharpness clearly indicates the correct upper bandgap as the solution (also marked by the orange band). The green band represents the best ML classifier model's classification of the defect energy level half bandgap location. The ML classifier correctly identify the half bandgap location and, in this case, aligns with the DPSS classification.

This analysis is repeated on the same defect parameters, adding random gaussian noise to the SRH lifetime curves. As shown in Fig. 6d, the noise is scaled up with decreasing excess carrier concentration to reproduce similar behaviour of actual lifetime measurements. The DPSS curves of the noisy data [Fig. 6e] shows that the ML regressor models are robust to the noise and continues to show accurate predictions of the defect parameters, with the best regressor predicting E_t within ± 0.004 eV and k within ±0.05. However, by adding noise to the SRH lifetime curves, the sharpness of the DPSS curves becomes ambiguous [Fig. 6f] and the DPSS approach now predicts the wrong half bandgap energy location. The ML still predicts the correct half bandgap energy location, showing the ML robustness to noise and presenting a case where the ML approach outperforms the DPSS. The comparison done in Fig. 6 is repeated 1,000 times for the same defect with random noise. The ML approach has an accuracy of 87.7% for half bandgap energy location prediction, while DPSS reaches 80.9% accuracy. When increasing the noise intensity, the accuracy of the DPSS drops to 44.7%, while the ML retains 76.4% accuracy. This is a strong demonstration that ML-based predictions are more reliable compared to DPSS-based predications when dealing with noisy signals.

The comparison between the DPSS and the ML has been repeated for 10,000 random defect parameters and set of SRH lifetime curves with random noise. In terms of classification, the DPSS approach predicts the correct half bandgap location 60% of the time, while the ML is correct 80% of the time over those 10,000 defects. Furthermore, the regression of the correct bandgap has shown an average defect energy level residual of 0.093 eV for the DPSS compared to almost of magnitude lower for the ML (0.011 eV). This is a clear indication that using the ML approach has a significant advantage over the DPSS approach in terms of its robustness to noise. This is quite remarkable considering the fact the ML classifier and regressors have been trained only on a dataset with no noise. We assume that the proven good robustness to potential measurement noise can be even further improved by training on a noisy dataset.

Experimental application

To be able to apply the ML approach to experimental measurements, an adaptation of the simulation parameters (such as wafer type, $N_{\rm dop}$, and temperature) to the desired measurement is necessary. Once the ML model is trained, the experimental measurements can be fed to the model in order to extract the defect parameters. This method is validated below by comparison to the traditional approach using the DPSS analysis.

For validation, an n-type Si wafer with $N_{\rm dop}=5.1\times10^{15}\,{\rm cm}^{-3}$ from Zhu et al. 13. is used. Lifetime measurements are done using TIDLS at eight different temperatures. The DPSS method is then used to extract $E_{\rm t}$ and k and the results are compared to ML predictions in Fig. 7. Each DPSS curve represents the best fit to the SRH model at that temperature. As discussed, the defects parameters are identified by a simultaneous minimization of the fit errors at all temperatures. The two solutions found by the DPSS method are marked by the black diamond symbol. The transparent ellipse represents the error-space defined by the error bars of this method. These two solutions are $(E_{\rm t}=-0.30\pm0.03\,{\rm eV};$ $k=1.87\pm1.18)$ for the lower half bandgap and $(E_{\rm t}=0.28\pm0.04\,{\rm eV};$ $k=1.01\pm0.44)$ for the upper half bandgap solution. The ML predictions are obtained by training an instance of different ML

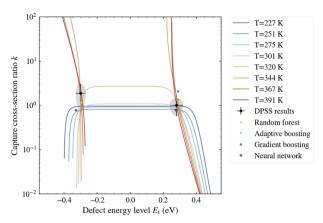


Fig. 7 Experimental measurements from machine learning predictions. DPSS curves of experimental measurements at multiple temperature. Black point and ellipse (error bar) show the traditional approach parameter extraction and coloured symbols the defect parameters extracted using machine learning.

training algorithms compared in this study: NN, RF and its boosting counterparts. Figure 7 shows that the ensemble models -RF, adaptive, and gradient boosting-all predict defect parameters within the DPSS solutions uncertainty range for both E_t and k, while NN predicts E_t within the DPPS solution uncertainty range but overshoot the value of k. Furthermore, calculating the sharpness of the DPSS curves' intersection points the half bandgap defect energy location in the lower half bandgap. In this experimental case, the ML approach concur with the DPSS results. The agreement between the ML models and the DPSS method highlights the success of this first-ever application of ML techniques for extraction of defect parameters. Based on this study, the approach can be extended to overcome the limitations of the currently used methods. For example, multi-level defects or multiple single-level defects modelling can be used to generate the training dataset and temperature-dependence of the defect parameters, such as the capture cross-sections, can be introduced into the physical models as well.

DISCUSSION

In this study, we proposed a ML-based approach to defect parameters extraction. By generating more than a million TIDLS lifetime curves with the SRH equation, ML models were trained and evaluated on simulated data. Overall, ensemble methods, such as RF, adaptive, and gradient boosting, seem to be the most promising algorithms to extract the defect parameters with an average R² score of 99%, whereas NN excels at predicting the half bandgap location of the defect with an average accuracy of 82.7%, outperforming the DPSS in predicting the half bandgap location of noisy SRH lifetime curves, which showcase an important advantage of the ML approach over the traditional approach. With no knowledge of the relationships predicated by the SRH equation, ML models provide insights regarding the physical limitations, such as when the lifetime curves are indistinguishable from one another, while still performing well in those conditions. Experimental measurements validated the ML approach in this successful defect parameter extraction using ML from lifetime measurements.

As more complex models are trained, additional insights can be gained that can contribute to a better understanding of defects. More complex models can also breach the limitations of traditional approaches as more modelling possibilities could be incorporated into the dataset generation step, such as multiple single-level defects and temperature-dependent capture

npj

cross-sections. The proposed approach can be generalized to any number of temperature measurements and wafer conditions and is not limited to Si wafers. The adaptability of this method ensures its uses in very varied scenarios and transfers its potential to other measurements techniques or material measured such as thin-film or perovskites with possibly more complex ML models.

METHODS

Dataset Simulation

The SRH^{18,19} equation is used to generate a dataset of simulated lifetime curves for a wide range of defect parameters ($E_{\rm tr}$ $\sigma_{\rm nr}$, $\sigma_{\rm p}$). The SRH lifetime $\tau_{\rm SRH}$ is given by:

$$\tau_{\text{SRH}} = \frac{\tau_{\text{p0}} \cdot (n_0 + n_1 + \Delta n) + \tau_{\text{n0}} \cdot (p_0 + p_1 + \Delta n)}{n_0 + p_0 + \Delta n} \tag{2}$$

with

$$\tau_{p0} := \left(\sigma_{p} \cdot \nu_{p} \cdot N_{t}\right)^{-1}; \tau_{n0} := \left(\sigma_{n} \cdot \nu_{n} \cdot N_{t}\right)^{-1} \tag{3}$$

$$n_1 := n_i \cdot e^{\frac{E_t}{k_b T}}; p_1 := n_i \cdot e^{\frac{-E_t}{k_b T}}$$

$$\tag{4}$$

In Eq. (2), n_0 (p_0) is the electron (hole) carrier concentration at thermal equilibrium and Δn is the excess carrier concentration. In Eq. (3), v_n (v_p) is the electron (hole) thermal velocity calculated using the model of Green et al.⁴⁰ and N_t is the defect density. In Eq. (4), $k_{\rm b}$ is Boltzmann's constant, T is the temperature and $n_{\rm i}$ is the intrinsic carrier concentration calculated using the model of Couderc et al.⁴¹, taking into account bandgap narrowing with the model of Yan et al.⁴². $E_{\rm t}$ is defined relative to the intrinsic energy of Si $E_{\rm i}$ ($E_{\rm t}=0$ eV is at mid-gap) and grows positive as $E_{\rm t}$ gets closer to the conduction band. At this stage, it is assumed that the defect parameters ($E_{\rm t}$, $\sigma_{\rm n}$, $\sigma_{\rm p}$) are independent of temperature.

For each combination of defect parameters (E_{ν} , $\sigma_{\rm n}$, $\sigma_{\rm p}$), 500 lifetime points are simulated as a feature vector of 100 Δn points at five different temperatures. Without loss of generality, p-type wafers are chosen with a bulk doping density $N_{\rm dop}$ fixed at 10^{15} cm⁻³, while $N_{\rm t}$ is fixed at 10^{12} cm⁻³. In total, 300,000 defect parameters combinations are chosen randomly within the reasonable physical range resulting in a million and a half-simulated lifetime curves ranging from 0.1 μ s to 10 ms. The ranges of the other simulation parameters are given in Table 1.

Machine learning training

The simulated dataset is randomly split into a training dataset (90% of the simulated data) used to train the ML model and a validation dataset (10%) used to evaluate the model. In the traditional defect parameter extraction approach, such as DPSS²⁰ and its derivatives 14,21,22 , two solutions for E_t are usually obtained, with one in the upper half of the bandgap (E_{t+}) and the second in the lower half of the bandgap (E_{t-}). In this study, for the regression of E_{t} , the training of the ML model is done separately for each half bandgap. The results are then combined and presented together. In order to identify in which half-bandgap the defect is located, the same approach is used to train a classifier model to predict the sign of E_t . The classifier outputs the probability of the defect being in the upper or lower half bandgap. The capture cross-section regression's target is the ratio k= $\sigma_{\rm n}/\sigma_{\rm p}$, which is usually reported in defect parameters' characterization $^{7.9,14,21,22}$. For the regression of k, the training of the ML model is done on the whole bandgap. In total, four ML models are trained to inverse the SRH equation: two regression models to predict E_{tr} one regression model to predict k; and one classification model to determine the defect energy location. The ML training flow is summarized in Fig. 8.

For each of the four ML models needed, different ML training algorithms can be used, defined in each grey box of Fig. 8. Random forests (RF)⁴³ are an ensemble of decision trees, where each tree is built from a random subset of the training data. The RF model outputs an average decision by all its decision trees. Enhancements of this method are possible through boosting, notably adaptive boosting^{37,44} and gradient boosting³⁸, which weights the decision of every tree to improve the predictive power of the overall model. A neural network (NN)^{36,45} is a layered group of nodes,

Table 1. Data simulation parameter range.			
Parameter	Symbol	Range or value	Unit
Defect energy level	E _t	[-0.55;0.55]	eV
Capture cross-sections	$\sigma_{\rm n},\sigma_{\rm p}$	$[10^{-17};10^{-13}]$	cm^2
Excess carrier concentration	Δn	[10 ¹³ ;10 ¹⁷]	${\rm cm}^{-3}$
Temperature	T	{200,250,300,350,400}	K
Defect density	N_{t}	10 ¹²	${\rm cm}^{-3}$
Doping density	N_{dop}	10 ¹⁵	${\rm cm}^{-3}$
Doping type	-	{p}	-

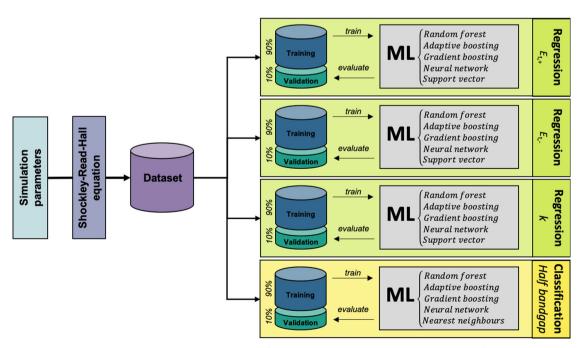


Fig. 8 Machine learning defect parameter extraction training flow. Dataset is simulated using random defect parameters within defined constraints and different machine learning models are trained for each of the prediction targets: $E_{t,+}$, $E_{t,-}$, k and the defect energy location.



called neurons, with a set of weights that are updated as it processes the training data. NN are widely used for different applications, such as image recognition⁴⁶, text translation⁴⁶ or beating Go world champions⁴⁷. Another widely used algorithm is support vector machines (SVM)^{48,49} which consists of mapping the training datapoints to a higher dimensional feature space where the data becomes easily separable. For classification problems, the K-nearest neighbours (KNN)⁵⁰ algorithm simply assigns the predicted label to be the majority vote of the closest known training data points. Figure 8 also shows that most ML training algorithms are used to train both regression and classification models, however, SVM is dedicated to regressions and KNN to classification.

In order to compare the different trained ML models, two scoring techniques are introduced. For the regression of defect parameters, predictions are made on the validation dataset and correlated to the true defect parameter values. To score the correlation, the coefficient of determination (R^2) is used S^3 :

$$R^{2} := 1 - \frac{\sum_{i=1}^{N} (y_{i}^{pred} - y_{i}^{true})^{2}}{\sum_{i=1}^{N} (y_{i}^{true} - y_{mean}^{true})^{2}};$$
 (5)

where 'pred' ('true') reference to the predicted (true) values of the validation dataset and y_i spans the N data points of the validation dataset. The R^2 score has a maximum value of 1 which indicates a perfect correlation and enables comparison between prediction targets. For the trained classification models, a confusion matrix is created, comparing predicted and true labels. The considered classification is binary, and predictions can have one of the following four outcomes: true positive (TP), true negative (TN), false negative (FN) and false positive (FP). The accuracy ('Acc') is a defined as 52 :

$$Acc := \frac{TP + TN}{TP + TN + FP + FN}$$
 (6)

The accuracy measures the ratio between correctly predicted points and the total number of points of the training dataset. The lifetime curves dataset simulation, ML model training and evaluation are implemented using Python and Scikit-learn⁵³. More information about the hyperparameter used can be found in the Supplementary Methods. The machine used for training has an Intel[®] Xeon[®] W-2145 processor, a central processing unit up to 3.70 GHz and a random-access memory of 64 GB. The training time is in the order of minutes for all the algorithms presented in this paper.

DATA AVAILABILITY

The authors declare that all data can be found on GitHub [https://github.com/ WhyBeU/DPML]. Any additional data can be accessed upon request.

CODE AVAILABILITY

The authors declare that all custom scripts and python code used for this study can be found on GitHub [https://github.com/WhyBeU/DPML].

Received: 10 March 2020; Accepted: 30 August 2020; Published online: 18 September 2020

REFERENCES

- 1. IPCC. Global warming of 1.5 °C. An IPCC special report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. In Masson-Delmotte, V. et al. (eds) (2018, In Press). https://archive.ipcc.ch/report/sr15/pdf/sr15_citation.pdf.
- 2. Green, M. A. Commercial progress and challenges for photovoltaics. *Nat. Energy* **1** 15015 (2016)
- 3. Needleman, D. B. et al. Economically sustainable scaling of photovoltaics to meet climate targets. *Energy Environ. Sci.* **9**, 2122–2129 (2016).
- Schmidt, J. et al. Impurity-related limitations of next-generation industrial silicon solar cells. In *IEEE Journal of Photovoltaics* vol. 3, pp. 114–118 (2013). https://doi. org/10.1109/JPHOTOV.2012.2210030.
- Coletti, G. Sensitivity of state-of-the-art and high efficiency crystalline silicon solar cells to metal impurities. *Prog. Photovolt.* 21, 1163–1170 (2013).

- Inglese, A., Lindroos, J., Vahlman, H. & Savin, H. Recombination activity of lightactivated copper defects in p-type silicon studied by injection- and temperaturedependent lifetime spectroscopy. J. Appl. 120, 125703 (2016).
- Sun, C., Rougieux, F. E., Degoulange, J., Einhaus, R. & Macdonald, D. Reassessment
 of the recombination properties of aluminium-oxygen complexes in n- and ptype Czochralski-grown silicon. *Phys. Status Solidi B* 253, 2079–2084 (2016).
- Schmidt, J. Temperature- and injection-dependent lifetime spectroscopy for the characterization of defect centers in semiconductors. *Appl. Phys. Lett.* 82, 2178–2180 (2003).
- Sun, C., Rougieux, F. E. & Macdonald, D. Reassessment of the recombination parameters of chromium in n- and p-type crystalline silicon and chromium-boron pairs in p-type crystalline silicon. J. Appl. 115, 214907 (2014).
- Schmidt, J., Krain, R., Bothe, K., Pensl, G. & Beljakowa, S. Recombination activity of interstitial chromium and chromium-boron pairs in silicon. J. Appl. 102, 123701 (2007).
- Birkholz, J. E., Bothe, K., Macdonald, D. & Schmidt, J. Electronic properties of ironboron pairs in crystalline silicon by temperature- and injection-level-dependent lifetime measurements. J. Appl. 97, 103708 (2005).
- Paudyal, B. B., McIntosh, K. R. & Macdonald, D. H. Temperature dependent electron and hole capture cross sections of iron-contaminated boron-doped silicon.
 In 2009 34th IEEE Photovoltaic Specialists Conference (PVSC), Philadelphia, PA, pp. 001588–001593 (2009). https://doi.org/10.1109/PVSC.2009.5411380.
- 13. Zhu, Y. et al. New insights into the thermally activated defects in n-type float-zone silicon. AIP Conf. Proc. 2147, 140014 (2019).
- Murphy, J. D., Bothe, K., Krain, R., Voronkov, V. V. & Falster, R. J. Parameterisation of injection-dependent lifetime measurements in semiconductors in terms of Shockley-Read-Hall statistics: an application to oxide precipitates in silicon. *J. Appl.* 111, 113709 (2012).
- Vargas, C. et al. Recombination parameters of lifetime-limiting carrier-induced defects in multicrystalline silicon for solar cells. Appl. Phys. Lett. 110, 092106 (2017).
- Diez, S., Rein, S., Roth, T. & Glunz, S. W. Cobalt related defect levels in silicon analyzed by temperature- and injection-dependent lifetime spectroscopy. J. Appl. 101. 033710 (2007).
- Rosenits, P., Roth, T., Glunz, S. W. & Beljakowa, S. Determining the defect parameters of the deep aluminum-related defect center in silicon. *Appl. Phys. Lett.* 91, 122109 (2007).
- Shockley, W. & Read, W. T. Statistics of the recombinations of holes and electrons. Phys. Rev. 87, 835–842 (1952).
- Hall, R. N. Electron-hole recombination in germanium. Phys. Rev. 87, 387–387 (1952).
- Rein, S. Lifetime Spectroscopy: A Method of Defect Characterization in Silicon for Photovoltaic Applications (Springer Science & Business Media, 2005).
- Zhu, Y., Gia, Q. T. L., Juhl, M. K., Coletti, G. & Hameiri, Z. Application of the Newton–Raphson method to lifetime spectroscopy for extraction of defect parameters. *IEEE J. Photovolt.* 7, 1092–1097 (2017).
- Morishige, A. E. et al. Lifetime spectroscopy investigation of light-Induced degradation in p-type multicrystalline silicon PERC. *IEEE J. Photovolt.* 6, 1466–1472 (2016).
- Peshek, T. J. et al. Insights into metastability of photovoltaic materials at the mesoscale through massive I–V analytics. J. Vac. Sci. Technol. B 34, 050801 (2016).
- Riley, D. & Johnson, J. Photovoltaic prognostics and heath management using learning algorithms. In Proc. 38th IEEE Photovoltaic Specialists Conference 001535–001539 (2012).
- Fada, J. S. et al. Electroluminescent image processing and cell degradation type classification via computer vision and statistical learning methodologies. In Proc. 44th IEEE Photovoltaic Specialist Conference 3456–3461 (2017).
- Haunschild, J. et al. Rating and sorting of mc-Si as-cut wafers in solar cell production using PL imaging. Sol. Energy Mater. Sol. C. 106, 71–75 (2012).
- Anwar, S. A. & Abdullah, M. Z. Micro-crack detection of multicrystalline solar cells featuring shape analysis and support vector machines. In *Proc. IEEE International* Conference on Control System, Computing and Engineering 143–148 (2012).
- Ni, B., Zou, P., Li, Q. & Chen, Y. Intelligent defect detection method of photovoltaic modules based on deep learning. Adv. Intel. Syst. Res. 161, 7 (2018).
- Ulaczyk, J., Morawiec, K., Zabierowski, P., Drobiazg, T. & Barreau, N. Finding relevant parameters for the thin-film photovoltaic cells production process with the application of data mining methods. *Mol. Inform.* 36, 1600161 (2017).
- 30. Brandt, R. E. et al. Rapid photovoltaic device characterization through bayesian parameter estimation. *Joule* **1**, 843–856 (2017).
- Sahu, H., Rao, W., Troisi, A. & Ma, H. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Adv. Energy Mater.* 8, 1801032 (2018).
- 32. Lu, S. et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **9**, 3405 (2018).
- Kurchin, R. C. et al. Semiconductor parameter extraction via current-voltage characterization and Bayesian inference methods. In Proc. 7th IEEE World Conference on Photovoltaic Energy Conversion 3271–3275 (IEEE, 2018).



- Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 3133–3181 (2014).
- 35. Nielsen, M. A. Neural Networks and Deep Learning (Determination Press, 2015).
- 36. Hinton, G. E. Connectionist learning procedures. Artif. Intell. 40, 185-234 (1989).
- Drucker, H. Improving regressors using boosting techniques. In Proc. International Conference on Machine Learning (1997).
- Friedman, J. H. Stochastic gradient boosting. Comput. Stat. Data. 38, 367–378 (2002).
- Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In Proc. 22nd International Conference on Machine Learning 625–632 (2005).
- Green, M. A. Intrinsic concentration, effective densities of states, and effective mass in silicon. J. Appl. 67, 2944–2954 (1990).
- 41. Couderc, R., Amara, M. & Lemiti, M. Reassessment of the intrinsic carrier density temperature dependence in crystalline silicon. J. Appl. 115, 093705 (2014).
- 42. Yan, D. & Cuevas, A. Empirical determination of the energy band gap narrowing in highly doped n+ silicon. *J. Appl.* **114**, 044508 (2013).
- 43. Breiman, L. Random forests. Mach. Learn. 45, 5-32 (2001).
- Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139 (1997).
- 45. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs] (2014).
- 46. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nat 521, 436-444 (2015).
- Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. Nat. 529, 484–489 (2016).
- Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers* 61–74 (MIT Press. 1999).
- 49. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. ACM T. Intel. Syst. Tec. 2, 1–27 (2011).
- Altman, N. S. An introduction to Kernel and nearest-neighbor nonparametric regression. Am. Stat. 46, 175–185 (1992).
- Glantz, S. A. & Slinker, B. K. Primer of Applied Regression and Analysis of Variance. (McGraw-Hill: New York, NY, 1990).
- Fawcett, T. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874 (2006)
- 53. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).

ACKNOWLEDGEMENTS

This work was supported by the Australian Government through the Australian Renewable Energy Agency [ARENA; Projects 2017/RND001 and 2017/RND017]. Yan Zhu acknowledges the support of the Australian Government through the Australian Centre for Advanced Photovoltaics (ACAP, project RG200768-G). The views expressed

herein are not necessarily the views of the Australian Government, and the Australian Government does not accept responsibility for any information or advice contained herein

AUTHOR CONTRIBUTIONS

Y.B. wrote the simulation and machine learning codes and the initial version of the manuscript. Z.H. supervised the work and contributed to the development of the methodology. Q.L. and J.D. contributed to the development of the methodology. Y.Z. contributed to the analysis of the experimental validation.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00410-7.

Correspondence and requests for materials should be addressed to Y.B.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.

© The Author(s) 2020

org/licenses/by/4.0/.