Extracting the parameters of two-energy-level defects in silicon wafers using machine learning

Sijin Wang, Priya Dwivedi, Yoann Buratti, Yan Zhu, Robert Lee Chin, and Ziv Hameiri

University of New South Wales, Sydney, NSW 2052, Australia
Email: sijin.wang@student.unsw.edu.au

With the alarming issues of climate change and energy shortage, the importance of sustainable energy technologies has significantly increased [1] [2]. Photovoltaic (PV) energy is one of the most promising renewable technologies [3], whose cost has dramatically decreased over the past thirty years [4]. According to the International Energy Agency (IEA), PV-generated electricity is the cheapest form of energy in human history [5]. Nevertheless, the price of PV systems should be further reduced to improve the accessibility of PV energy in developing countries [6] and to obtain the full potential of this technology. One of the key methods to do this is by improving the power-conversion efficiency of solar cells [7]. A key limitation towards improving cell efficiencies is the recombination-active defects in the silicon bulk [8]. To minimise the detrimental effects of these defects, it is essential to predict their energy levels (E_t) and their electron and hole capture cross-sections (σ_n and σ_p, respectively) [9] [10].

Commonly, bulk defects are classified as single-energy-level [9] or multi-energy-level [11]. This study will focus on two-level defects, which are a subset of multi-energy-level defects. Temperature- and injection-dependent lifetime spectroscopy (TIDLS) [12] is often used to extract the parameters of the defect [10] [13] [12]. In the case of single-energy-level defects, a few analysis methods are commonly employed. The most common method is the defect parameter solution surface (DPSS) method [12]. Other methods include the linearised DPSS method [14], the Newton Raphson method [10], and the machine learning (ML)-based approach [15]. However, in the case of two-level defects, there is only the pioneering study by Zhu *et al.* [13]. Here, we present a novel ML-based approach to extracting the defect parameters of two-level defects.

To train the developed ML algorithms, a dataset of 800,000 defects and the corresponding lifetime curves with different temperatures was generated using the Sah-Shockley equation [11] with random selections of the defect parameters (E_{t1}, E_{t2}, σ_{n1}, σ_{p1}, σ_{n2}, σ_{p2}, and N_t; the subscript '1' or '2' indicates the level of the defect):

$$\tau_{\mathrm{two-level}} = \frac{1+\frac{\sigma_{n1}v_n n_1+\sigma_{p1}v_p p}{\sigma_{p1}v_p p_1+\sigma_{n1}v_n n}+\frac{\sigma_{p2}v_p p_2+\sigma_{n2}v_n n}{\sigma_{n2}v_n n_2+\sigma_{p2}v_p p}}{N_t(n_0+p_0+\Delta n)[\left(\frac{\sigma_{n1}\sigma_{p1}v_n v_p}{\sigma_{p1}v_p p_1+\sigma_{n1}v_n n}\right)+\left(\frac{\sigma_{n2}\sigma_{p2}v_n v_p}{\sigma_{n2}v_n n_2+\sigma_{p2}v_p p}\right)]}, \quad \text{Eq. 1}$$

n_1, n_2, p_1, and p_2 are defined by Eqs. 2-5:

$$n_1 = n_i e^{\frac{E_{t1}-E_i}{k_b T}}, \quad \text{Eq. 2}$$

$$n_2 = n_i e^{\frac{E_{t2}-E_i}{k_b T}}, \quad \text{Eq. 3}$$

$$p_1 = n_i e^{\frac{-E_{t1}+E_i}{k_b T}}, \quad \text{Eq. 4}$$

$$p_2 = n_i e^{\frac{-E_{t2}+E_i}{k_b T}}. \quad \text{Eq. 5}$$

In Eq. 1, $\tau_{\text{two-level}}$ refers to the recombination lifetime of the two-level defect, σ_{n1} (σ_{p1}) is the electron (hole) capture cross-section of the first energy level, σ_{n2} (σ_{p2}) is the electron (hole) capture cross-section of the second energy level, and v_n (v_p) is the electron (hole) thermal velocity calculated using the model of Green *et al.* [16]. Furthermore, n (p) is the electron (hole) density, N_t is the total defect concentration, n_0 (p_0) is the thermal equilibrium electron (hole) density, and Δn is the excess carrier density. In Eqs. 2 to 5, n_i is the intrinsic carrier density calculated using the model of Couderc *et al.* [17] with bandgap narrowing by the model of Yan *et al.* [18], E_{t1} and E_{t2} are the defect energy levels, E_{t1} is the transition energy between the most positively charged state and the middle charge state while E_{t2} is the transition energy between the most negatively charged state and the middle charge state [13] [19]. Once the dataset was created, it was randomly split into a training set (90% of the simulated data) and a validation set (10%). E_i is the intrinsic fermi energy of silicon, k_b is the Boltzmann's constant, and T is the bulk temperature.

Two solutions for E_t are usually obtained in one-level defect parameter extraction methods such as DPSS [12], one in each bandgap half (higher or lower than E_i). As a result, these two cases require separate training instances for the ML-based one-level defect parameter extraction [15]. Similarly, for a two-level defect, the DPSS residual map using TIDLS data proposed by Yan [13] identifies multiple possible solutions in different bandgap halves as well. Since two-level-defects have two energy levels, this study will use four cases: both energy levels above E_i (named here Set 11); both below E_i (Set 00); $E_{t1} > E_i$ and $E_{t2} < E_i$ (Set 10); $E_{t1} < E_i$ and $E_{t2} > E_i$ (Set 01).

The proposed method is divided into two steps as shown in Fig. 1. In the first step, the defect parameters are extracted by regression for the four cases listed above. In the second step, the ML algorithm is trained to classify the most likely solution out of the four cases.

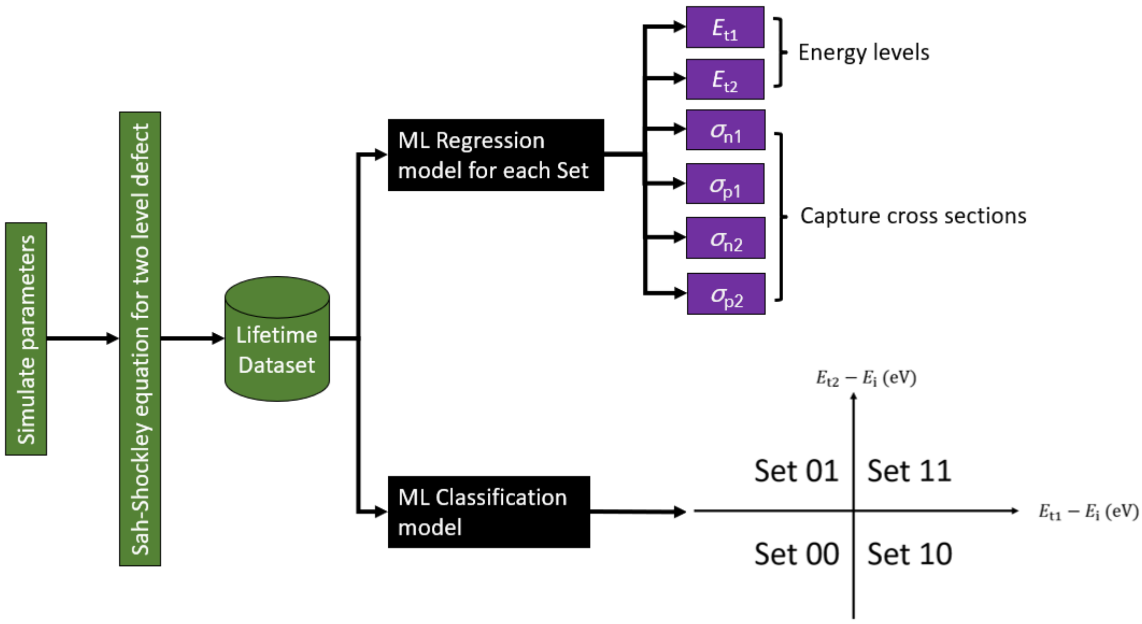

Figure 1: A flow chart for defect parameter extraction using machine learning.

To evaluate the performance of the developed ML models, two scoring methods are used. For the regression task, the coefficient of determination (R^2 score) is used as defined by Eq. 6 [20]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i^{\text{pred}} - y_i^{\text{true}}\right)^2}{\sum_{i=1}^{N}(y_i^{\text{true}} - y_{\text{mean}}^{\text{true}})^2}, \qquad \text{Eq. 6}$$

where the label 'pred' and 'true' refers to the predicted and true values of the test dataset respectively and R^2 ranges from zero (very poor prediction) to unity (perfect prediction). The accuracy matrix is utilised for the classification task, which is the ratio of the number of correct predictions to the total number of predictions [21].

For the regression task, a random forest model [22] was trained. The predicted vs simulated values of the validation set for a *p*-type wafer (bulk doping of 10^{15} cm^{-3}, Set 11) are presented in Fig. 2, the red line represents the ideal prediction while the green dots are the actual prediction. The defect parameters of the first level (E_{t1}, σ_{n1}, σ_{p1}) are extracted with a high level of accuracy ($R^2 > 0.9$). The predictions of E_{t2}, σ_{n2}, σ_{p2}, are reasonable, but less accurate ($R^2 > 0.6$). The difference in the prediction qualities could be as the Sah-Shockley recombination lifetime is more sensitive to changes in the first-level defect parameters in *p*-type wafers.

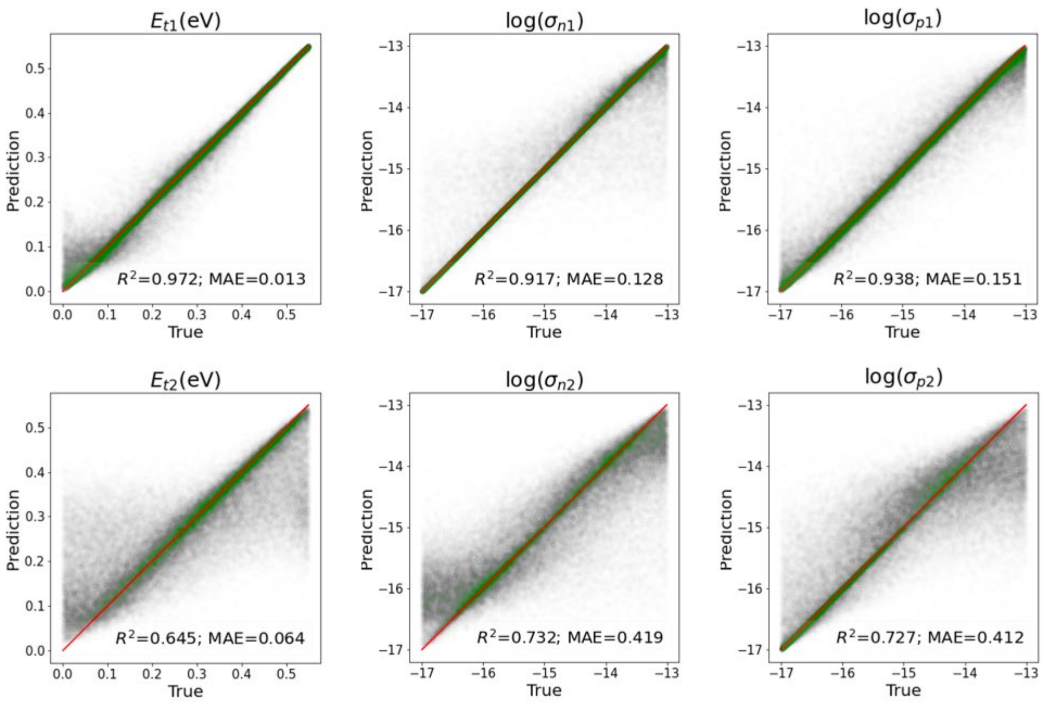

Figure 2: The predicted vs simulated parameters for a *p*-type wafer in Set 11 (bulk doping density of 10^{15} cm^{-3}).

The correlation coefficients of the different sets are summarised in Table 1. The reason for the use of the logarithm of the capture cross sections is that it ranges over several orders of magnitude (typically 10^{-13} cm^{-3} to 10^{-17} cm^{-3} [23]). It can be seen that for *p*-type doping when the set number begins with the subscript "1", meaning $E_{t1} > E_i$, the prediction for E_{t1}, σ_{n1}, and σ_{p1} are more accurate, and vice versa for E_{t2}. The reason for this pattern is still under investigation.

Table 1: The coefficient of correlation (R^2 score) for the four sets (a p-type wafer)

	E_{t1}	$\log(\sigma_{n1})$	$\log(\sigma_{p1})$	E_{t2}	$\log(\sigma_{n2})$	$\log(\sigma_{p2})$
Set 11	0.972	0.919	0.939	0.643	0.726	0.726
Set 10	0.936	0.987	0.941	0.821	0.807	0.703
Set 01	0.813	0.916	0.726	0.693	0.689	0.783
Set 00	0.827	0.882	0.739	0.934	0.849	0.766

The confusion matrix of the classification model is shown in Table 2, the rows correspond to the true class while the columns represent the predicted class; the number represents the percentage of this prediction compared to the dataset size. The correct predictions are shown in the diagonal. At this stage, the average accuracy for the four sets is about 67%. Furthermore, the specificity of each set is shown on the rightest column, which is the ratio between the correct classification over the actual population of each set. It seems that the ML model has difficulties identifying Set 11 correctly.

Table 2: The confusion matrix of the classification model.

True \ Prediction	Set 11	Set 10	Set 01	Set 00	Specificity
Set 11	13.1%	5.9%	0.8%	5.3%	52.2%
Set 10	4.5%	17.9%	1.6%	1.2%	71.0%
Set 01	0.6%	1.7%	18.4%	4.5%	73.0%
Set 00	2.9%	0.3%	3.3%	18.2%	73.7%

To summarise, this research presented a novel ML-based approach to solving a challenging problem – the extraction of the defect parameters of a two-level defect from TILDS data. The defect parameters of the four possible cases are first extracted. A classification model then predicts the most likely solution. We note a difference between the prediction accuracy of E_{t1}, σ_{n1}, σ_{p1} and E_{t2}, σ_{n2}, σ_{p2}. This is due to the difference in the sensitivity of the lifetime with different parameters for a given set of doping densities and/or temperatures. However, this is not a limitation when the same two-level defect is present in both *p*- and *n*-type wafers. Future work will focus on improving prediction accuracy and uncovering the reason for better predictions of some parameters than others.

Reference

[1] S. Stec and B. Baraj, Eds., *Energy and Environmental Challenges to Security*, 1st ed. Dordrecht, Netherlands: Springer: In cooperation with NATO Public Diplomacy Division, 2009.

[2] B. Verlie, *Learning to Live with Climate Change: From Anxiety to Transformation*, 1st ed. New York, USA: Routledge, 2021.

[3] Q. Khalid, J. Ikram, and N. Arshad, “A collaborative approach to operate high powered devices on small-scale PV systems,” *Energy Procedia*, vol. 111, pp. 895–903, Mar. 2017, doi: 10.1016/j.egypro.2017.03.252.

[4] A. Boretti and S. Castelletto, “Cost and performance of CSP and PV plants of capacity above 100 MW operating in the United States of America,” *Renew. Energy Focus*, vol. 39, pp. 90–98, Dec. 2021, doi: 10.1016/j.ref.2021.07.006.

[5] C. B. Staff, “Solar is now ‘cheapest electricity in history’, confirms IEA,” *Carbon Brief*, Oct. 13, 2020. https://www.carbonbrief.org/solar-is-now-cheapest-electricity-in-history-confirms-iea/ (accessed Aug. 28, 2022).

[6] Z. Dobrotkova, K. Surana, and P. Audinet, “The price of solar energy: comparing competitive auctions for utility-scale solar PV in developing countries,” *Energy Policy*, vol. 118, pp. 133–148, Apr. 2018, doi: 10.1016/j.enpol.2018.03.036.

[7] C. Messmer *et al.*, “The race for the best silicon bottom cell: efficiency and cost evaluation of perovskite–silicon tandem solar cells,” *Prog Photovolt*, vol. 29, no. 7, pp. 744–759, Nov. 2021, doi: 10.1002/pip.3372.

[8] J. Schmidt *et al.*, “Impurity-related limitations of next-generation industrial silicon solar cells,” *IEEE J. Photovolt.*, vol. 3, no. 1, pp. 114–118, Jan. 2013, doi: 10.1109/JPHOTOV.2012.2210030.

[9] W. Shockley and W. T. Read, “Statistics of the recombinations of holes and electrons,” *Phys. Rev.*, vol. 87, no. 5, pp. 835–842, Apr. 1952, doi: 10.1103/PhysRev.87.835.

[10] Y. Zhu, *Advanced Characterization of Defects in Silicon Wafers and Solar Cells*, 1st ed. Sydney, NSW, Australia: UNSW PHD thesis, 2019.

[11] C.-T. Sah and W. Shockley, “Electron-hole recombination statistics in semiconductors through flaws with many charge conditions,” *Phys. Rev.*, vol. 109, no. 4, pp. 1103–1115, Jul. 1958, doi: 10.1103/PhysRev.109.1103.

[12] S. Rein, *Lifetime Spectroscopy: A Method of Defect Characterization in Silicon for Photovoltaic Applications*, 1st ed. Berlin, Heidelberg, New York, USA: Springer, 2005.

[13] Y. Zhu, C. Sun, T. Niewelt, G. Coletti, and Z. Hameiri, “Investigation of two-level defects in injection dependent lifetime spectroscopy,” *Sol. Energy Mater Sol. Cells*, vol. 216, p. 110692, Jul. 2020, doi: 10.1016/j.solmat.2020.110692.

[14] J. D. Murphy, K. Bothe, R. Krain, V. V. Voronkov, and R. J. Falster, “Parameterisation of injection-dependent lifetime measurements in semiconductors in terms of Shockley-Read-Hall statistics: an application to oxide precipitates in silicon,” *J. Appl. Phys.*, vol. 111, no. 11, p. 113709, Jun. 2012, doi: 10.1063/1.4725475.

[15] Y. Buratti, Q. T. Le Gia, J. Dick, Y. Zhu, and Z. Hameiri, “Extracting bulk defect parameters in silicon wafers using machine learning models,” *npj Comput. Mater*, vol. 6, no. 1, p. 142, Sep. 2020, doi: 10.1038/s41524-020-00410-7.

[16] M. A. Green, “Intrinsic concentration, effective densities of states, and effective mass in silicon,” *J. Appl. Phys.*, vol. 67, no. 6, pp. 2944–2954, Mar. 1990, doi: 10.1063/1.345414.

[17] R. Couderc, M. Amara, and M. Lemiti, “Reassessment of the intrinsic carrier density temperature dependence in crystalline silicon,” *J. Appl. Phys.*, vol. 115, no. 9, p. 093705, Mar. 2014, doi: 10.1063/1.4867776.

[18] D. Yan and A. Cuevas, “Empirical determination of the energy band gap narrowing in highly doped n $^+$ silicon,” *J. Appl. Phys.*, vol. 114, no. 4, p. 044508, Jul. 2013, doi: 10.1063/1.4816694.

[19] F. E. Rougieux, C. Sun, and D. Macdonald, “Determining the charge states and capture mechanisms of defects in silicon through accurate recombination analyses: a review,” *Sol. Energy Mater Sol. Cells*, vol. 187, pp. 263–272, Jul. 2018, doi: 10.1016/j.solmat.2018.07.029.

[20] S. Glantz and B. Slinker, *Primer of Applied Regression & Analysis of Variance*, 1st ed. New York, USA: McGraw-Hill Education, 2000.

[21] E. Alpaydin, *Machine learning*, 2nd ed. Cambridge, UK; Massachusetts, USA: The MIT Press, 2021.

[22] A. C. Müller, *Introduction to Machine Learning with Python: a Guide for Data Scientists*, 1st ed. Sebastopol, California, USA: O’Reilly Media, Inc, 2017.

[23] E. H. Nicollian and J. R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology*. Wiley, 1982.