IOP Publishing Semicond. Sci. Technol. 41 (2026) 015021 https://doi.org/10.1088/1361-6641/ae3850

Semiconductor Science and Technology

PAPER

RECEIVED
12 October 2025

REVISED
18 December 2025

ACCEPTED FOR PUBLICATION
14 January 2026

PUBLISHED
23 January 2026

Computer vision-based method for quantifying iron-related defects in silicon solar cells

Oleg Olikh*, Oleksii Zavhorodnii and Yulia Perets

Taras Shevchenko National University of Kyiv, Kyiv 01601, Ukraine
* Author to whom any correspondence should be addressed.

E-mail: olegolikh@knu.ua

Keywords: defect, Si solar cell, iron contamination, machine learning, computer vision

Supplementary material for this article is available online

Abstract

This study demonstrates the feasibility of employing transfer learning from pre-trained computer vision (CV) models to predict the iron concentration in silicon solar cells, even when the available training datasets are extremely limited. The predictions were based on the kinetic dependencies of the short-circuit current following FeB pair dissociation, which was converted into images using wavelet transformation. The performance of various combinations of CV models and regression algorithms was systematically analyzed. Specifically, several state-of-the-art CV architectures, including EfficientNetB7, MobileNetV2, NASNetLarge, ResNet152V2, Xception, and YOLOv4, were utilized either as classifiers or as feature extractors. Regression models, namely random forest, gradient boosting (GB), eXtreme GB, support vector regression (SVR), and deep neural networks (DNNs), were trained to predict iron concentration from the extracted features. Training and testing were performed using both simulated and experimental datasets. In both cases, EfficientNetB7 and NASNetLarge provided the most informative features for subsequent regression. Among the regression algorithms, SVR and DNNs were identified as the most effective. These models achieved mean squared error, mean absolute percentage error, median absolute percentage error, and R^2 values of up to 0.001, 6%, 4%, and 0.999, respectively, for the simulated data, and 0.008, 10%, 5%, and 0.996 for the experimental data.

1. Introduction

Owing to the urgent need to address environmental challenges and growing global demand for renewable energy, the deployment of photovoltaic (PV) systems has been rapidly increasing worldwide. In particular, solar PV generation exceeds 1600 TWh in 2023 [1, 2], rising by approximately 30% in 2024 [3], and forecasts indicate that the total installed capacity will surpass 6 TW by 2030 [1]. Meanwhile, crystalline silicon PVs, which have benefited from decades of scientific advancement and continuous cost reductions, continued to dominate the market in 2024, accounting for approximately 98% of the global share [4, 5].

As in other semiconductor devices, defects play a decisive role in determining the operating parameters of the solar cells. Therefore, diagnosing defects, particularly determining their concentrations, is critical for maintaining stable performance of PV systems. In recent years, researchers have increasingly complemented established defect characterization methods with machine learning (ML) approaches that improve the accuracy, speed, and cost efficiency of these analyses. The use of ML methods for analyzing macroscopic defects (such as cracks, finger failures, hotspots, and scratches) and point defects, however, differs significantly. Researchers typically detect macroscopic defects in PV systems using two main approaches [6, 7]. The first one, electrical testing techniques, involves analyzing characteristic electrical curves of parameters such as current, voltage, and power. The second approach, imaging-based techniques, involves analyzing electroluminescence (EL) [8] or photoluminescence [9] images of solar cells. Numerous review studies demonstrate extensive use of ML in both approaches [10–16].

Point defects represent a significant limiting factor in the performance of PV devices; however, the development of ML methodologies specifically tailored for their analysis remains comparatively limited. Existing applications of ML in microscopic defect characterization can be broadly categorized into several distinct approaches. One such approach focuses on enhancing conventional defect-analysis techniques through the integration of Artificial Intelligence methods for processing and interpreting the resulting experimental signals.

For example, Buratti *et al* [17] employed regression algorithms, including random forest (RF), gradient boosting (GB), and deep neural networks (DNN), to analyze dependencies derived from temperature- and injection-dependent lifetime spectroscopy (TIDLS). They trained the models on more than one million simulated curves, which enabled accurate estimation of silicon defect energy levels and carrier capture cross-sections. In addition, unlike the conventional fitting of signals with the Shockley–Read–Hall equation, their approach can also predict the energy level position at half of the bandgap. An extension of this approach was presented in [18], where the methodology incorporated a convolutional neural network (CNN) to analyze images derived from a family of lifetime curves measured at different temperatures, in addition to applying a RF model to the standard TIDLS signal. In that study, the CNN was used both to perform the classification of the half-bandgap position of the energy level and to extract features, which were subsequently used by the RF. As in the earlier work, the models were trained on a dataset consisting of several hundred thousand synthetic samples. An alternative TIDLS signal processing strategy was also investigated by Wang *et al* [19], who used CNNs to analyze one-dimensional signals and thereby extract parameters associated with two-energy-level defects in silicon. ML methods are also employed for the analysis of Raman spectra [20]. In this study, the spectra of electron-irradiated GaAs were examined using linear discriminant analysis models. These models were trained on 6000 experimentally acquired spectra, enabling the identification of radiation-induced defects.

An alternative approach is based on determining defect parameters by analyzing the characteristics of devices, primarily solar cells, that are directly affected by such defects. For silicon solar cells, for example, a methodology has been proposed to estimate the concentration of contaminant impurities from the magnitude of the ideality factor obtained from current–voltage (I–V) characteristics [21] or from variations in PV conversion parameters [22]. In both studies, classical regression algorithms (DNN, RF, support vector regression (SVR), and GB) were employed. The numerical values of parameters extracted from the I–V characteristics served as input features, and the models were trained on tens of thousands of current–voltage curves simulated under different defect parameters. A closely related approach was presented by Haidari *et al* [23], who used thirteen parameters extracted from the I–V curves of CIGS solar cells as inputs to a DNN to predict the spatial distribution and concentration of six bulk and surface defects. The inverse problem, namely the determination of PV conversion parameters based on predefined defect concentrations, was examined by Kim *et al* [24] for perovskite solar cells. In that study, RF, XGBoost, linear regression, and multilayer perceptron algorithms were evaluated, and the RF model delivered the highest performance. In all four studies mentioned above, the SCAPS-1D simulation tool was consistently used to generate the synthetic I–V characteristics that formed the training datasets.

Beyond the detection and characterization of defects in actual devices, a distinct research direction focuses on accelerating and improving density functional theory (DFT) and molecular dynamics (MD) calculations of defect parameters. For example, several studies have demonstrated the use of graph neural networks, trained on DFT-calculated data, to estimate vacancy formation energies [25, 26] and to evaluate the electronic structure of charged defects in GaAs [27]. Graph convolutional networks have also been applied to MD-generated datasets for predicting vacancy diffusion paths in high-entropy alloys [28] and intrinsic defects in perovskites [29]. In addition, DFT datasets have been integrated with ML methods to identify formation enthalpies and ionization energies of impurity defects [30] and to determine the equilibrium configurations of defects in emerging materials [31].

Nevertheless, relatively few ML methodologies are currently available for the characterization of point defects, especially regarding the experimental determination of their parameters. One of the main challenges in applying ML methods effectively is that training the models requires a large amount of labeled data [12]. In practice, researchers often cannot obtain such large volumes of experimental data; therefore, they commonly employ approaches such as simulations, in which hundreds of thousands of dependencies are computed [17–19, 22, 32]; physics-informed neural networks (PINNs), which incorporate physical laws into the loss function to generate synthetic data [33, 34]; or transfer learning, in which a model trained on one task is adapted to another related task [35, 36]. However, simulations can be highly demanding in terms of time and computational resources; PINNs are primarily suitable for phenomena described by partial differential equations, and pre-trained models are not available for all types of physical problems. At the same time, one of the most extensively studied tasks in ML is computer vision (CV), for which many pre-trained models have been publicly released. Moreover, these models

are typically trained on extremely large standard datasets. For example, EfficientNetB7 was trained on approximately 1.2 million images from the ImageNet dataset.

This study primarily aimed to apply standard pre-trained CV models to analyze the electrophysical measurement results related to point defects. In particular, we focus on quantifying iron in boron-doped crystalline silicon solar cells by examining short-circuit current (I_{SC}) relaxation following intense illumination. Iron is among the dominant metallic contaminants that degrade the efficiency of these structures. Although current research has increasingly focused on next-generation solar cells, particularly perovskite-based devices, silicon structures [37, 38], as previously discussed, constitute the core of the PV market, highlighting the continued relevance of the present work. In Si:B, iron readily associates with boron to form FeB pairs, and these complexes can be dissociated by strong illumination [39, 40]. In fact, the aforementioned I_{SC} variations directly reflect the recovery process of iron–boron pairs [41].

It should be noted that the use of well-established CV benchmark architectures such as YOLO, MobileNetV2, EfficientNet, ResNet, Xception, GoogleNet, and other CNNs is a common approach for identifying macrodefects from EL measurements [6, 8, 42–47]. However, in this case, the measurement result is an image, makes the approach relatively straightforward. In our case, it was necessary to transform the time dependence into an image representation. Standard approaches to solving such problems involve the use of Fourier or wavelet transforms, and the latter were applied in this study. In PVs, wavelet transforms are typically used for processing solar cell images to enhance the detection of macrodefects [48, 49], but they can also be employed to convert one-dimensional non-stationary signals into two-dimensional spectrograms and thereby enable the effective extraction of subtle features [50].

By applying CV models to wavelet spectrograms represented as images, we generated high-dimensional feature vectors and used them as inputs for traditional regression models. The $I_{SC}(t)$ dependencies for training the regression models were obtained via simulations and experimental measurements. In both cases, the hybrid transfer learning approach produced predictions with sufficiently high accuracy (within a few percent) even when training models on small datasets containing fewer than 30 samples. Importantly, the proposed approach is highly versatile and can be extended to a wide range of tasks related to defect characterization and other applications.

To summarize the novelty and contribution of our work, as well as its distinction from previous studies on semiconductor defects, we note the following. The use of CNNs for analyzing defect-related electrophysical dependencies has been explored previously [18]. In that study, however, image construction required a set of curves measured under different conditions, specifically at various temperatures, and the model was trained from scratch, which demanded a very large training dataset. In contrast, our approach relies on a single kinetic dependency and leverages the capabilities of pre-trained CV models. Standard CV models for defect detection in solar cells have also been reported [6, 8, 45–47], although prior work focused on macro-defects and processed naturally acquired images from conventional cameras. In our case, the emphasis is on point-defect characteristics, and the images used as input are generated from electrophysical measurements. In the analysis of solar cells, wavelet transforms have been applied to one-dimensional dependencies [50] in addition to their use in improving defect detection in photographic images [48]. In those studies, however, the resulting wavelet coefficients were used as features in regression algorithms rather than for constructing images, which is the approach adopted in the present work. More broadly, to the best of our knowledge, one-dimensional signal-to-image conversion for CNN input preparation has typically been achieved either by employing a set of curves [18] or by digitizing standard graphs produced in software such as Origin [51]. The use of the wavelet transform as a preprocessing step for CNNs is therefore novel. Finally, our methodology is designed to function effectively with extremely small datasets, which facilitates the practical application of the proposed approach.

2. Methodology

2.1. General outline of the method

Figure 1 illustrates the workflow of the ML pipeline used to extract iron contamination from $I_{SC}(t)$ dependencies. The process consisted of three main blocks: data acquisition, CNN feature processing, and predictive regression. The first stage involves either simulating or experimentally measuring the time dependence of the short-circuit current in a solar cell after the induced decay of the FeB pairs. These procedures are described in detail in sections 2.2 and 2.3. For the experimental curves, the data were smoothed using a Savitzky–Golay filter [52]. Subsequently, a continuous wavelet transform (CWT) [53] was applied to convert the one-dimensional time dependencies into two-dimensional spectrograms represented as images, where each point corresponds to the amplitude of the wavelet coefficient at a specific time and frequency. The Morlet wavelet was employed, and the procedure was implemented using the

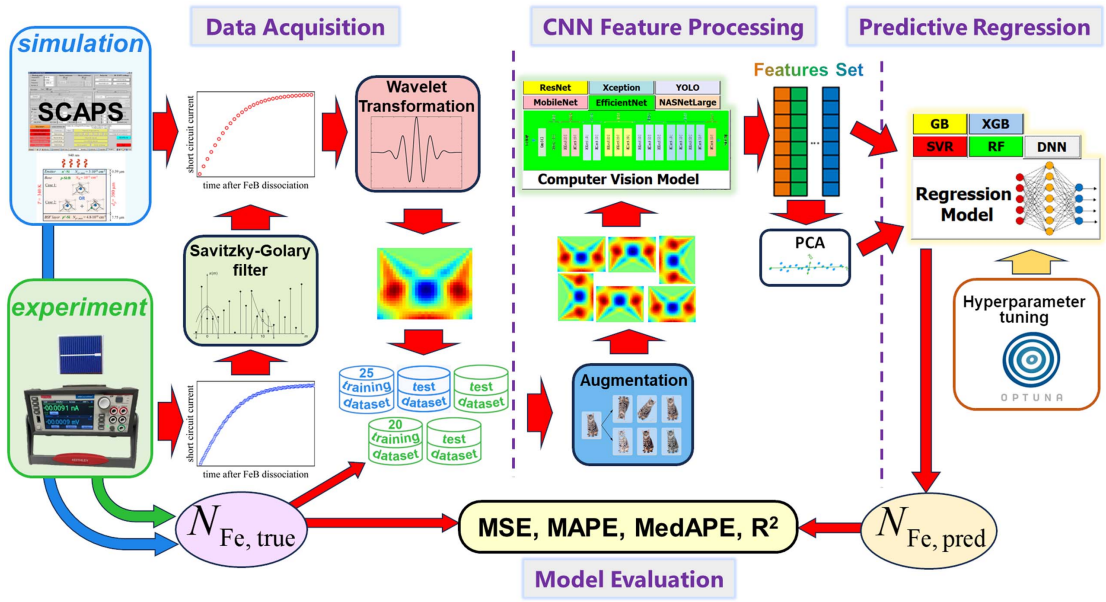

Figure 1. Workflow of the ML pipeline.

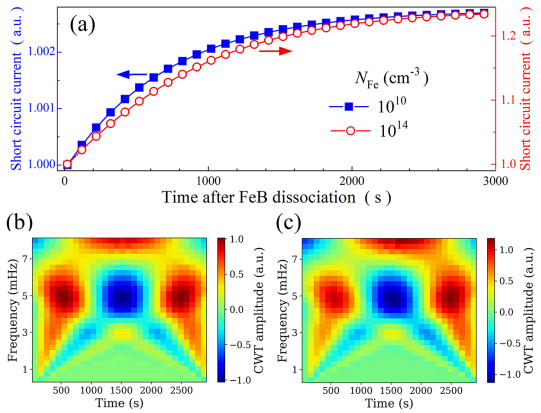

Figure 2. Simulated time dependencies of short-circuit current (a) and corresponding wavelet spectrograms for iron concentrations of 10^{10} cm^{-3} (b) and 10^{14} cm^{-3} (c). The data in panel a are shown with filled squares for the concentration corresponding to panel b and with open circles for that corresponding to panel (c).

Python package PyWavelets. Examples of the resulting images are shown in figures 2(b) and (c). Data augmentation was then performed by flipping the images along the *x*- and *y*-axes and rotating them by 90°, 180°, and 270°. This procedure is known to improve the accuracy of ML model predictions, particularly when only small datasets are available [54].

During the CNN feature processing stage, all images (both original and augmented) were processed using one of the standard CV models to extract a feature set for each image. The selected models and feature extraction settings are described in section 2.4. No CNN fine-tuning was performed, and the models were used in their pre-trained form, as downloaded. In general, the dimensionality of the feature vectors obtained from the CNN outputs substantially exceeds the number of available samples, implying a high degree of redundancy. Therefore, to enable comparison and mitigate this effect, principal component analysis (PCA) was applied in some cases to reduce the feature dimensionality with negligible loss of total variance.

The obtained feature sets served as inputs to the regression models based on one of the standard algorithms described in section 2.5, which aimed to predict the iron concentration (N_{Fe}) in the solar cell. In the first case, the regression models were trained on a simulated training dataset and tested on both the simulated test dataset and experimental data. In the second case, a portion of the experimental results was used for training, whereas the remaining part was reserved for testing the corresponding

models. During training, feature sets derived from the original wavelet spectrograms and their augmented versions were treated as separate samples. During testing, the median of the predicted values obtained from the original and augmented images was used as the final prediction. Model performance was evaluated using the metrics described in section 2.6.

2.2. Simulation details

The time-dependent short-circuit current, $I_{SC}(t)$, was determined by simulating the I–V characteristics of a silicon n^+–p–p^+ structure under monochromatic illumination using SCAPS-1D.3.3.11. SCAPS-1D [55] is a widely adopted software for solar cell modeling that incorporates the effects of defect states [56–61].

During the simulations, the base thickness of the structure was set to 380 μm, and boron was used as the doping element with a concentration of $N_B = 1.36 \times 10^{15}$ cm^{-3}. The temperature was maintained at 340 K, and monochromatic illumination with a wavelength of 940 nm and an intensity of 5 W m^{-2} was applied, corresponding to the experimental conditions (see section 2.3). One of the modeling parameters was the total concentration of iron impurity atoms, N_{Fe}. It was assumed that Fe atoms were uniformly distributed throughout the base and p^+ layer of the solar cell and could exist either in interstitial positions, with a concentration N_{Fe_i}, or as FeB pairs, with a concentration N_{FeB}. The time dependence of N_{Fe_i} after pair dissociation follows the well-known expression [62, 63]:

$$N_{Fe_i}(t) = \left(N_{Fe_i,0} - N_{Fe_i,eq}\right) \times \exp\left(-t/\tau_{ass}\right) + N_{Fe_i,eq}, \tag{1}$$

where $N_{Fe_i,0}$ is the concentration of interstitial iron atoms formed due to FeB pair dissociation, $N_{Fe_i,0} = N_{Fe_i}(t=0) = N_{Fe}$; $N_{Fe_i,eq}$ is the portion of interstitial iron atoms that remain unpaired in the equilibrium state $N_{Fe_i,eq} = N_{Fe_i}(t \to \infty)$, according to [62, 63]

$$N_{Fe_i,eq} = \frac{N_{Fe}}{\left[1 + N_B \cdot A_z \cdot \exp\left(\frac{E_b}{kT}\right)\right]\left[1 + \exp\left(\frac{E_F - E_{Fe_i}}{kT}\right)\right]}, \tag{2}$$

E_b is the binding energy of the FeB pairs (taken as 0.582 eV [63]), A_z depends on the number of possible orientations of the pair and lattice site density (taken as 10^{-23} cm^3 [62]), E_F is the Fermi level, E_{Fe_i} is the position of the donor Fe$_i$ level relative to the valence band maximum (taken as 0.394 eV [40]), τ_{ass} is the characteristic time of the complex association, according to [40, 64, 65]

$$\tau_{ass} = A \times \frac{T}{N_A} \exp\left(\frac{E_m}{kT}\right), \tag{3}$$

E_m is the energy of Fe_i^+ migration (taken as 0.66 eV [40, 64, 65]), A is the pre-exponential constant (taken as $5.7 \times 10^5 \frac{s}{K\,cm^3}$ [65]). The iron–boron pair concentration N_{FeB} was estimated as follows:

$$N_{FeB}(t) + N_{Fe_i}(t) = N_{Fe}. \tag{4}$$

Overall, the concentrations of iron-related defects depended not only on time but also on their spatial position within the structure, reflecting the non-uniformity of the $(E_F$–$E_{Fe_i})$ difference.

The parameters used in the simulations are summarized in table 1. Additional details on the procedure for modeling the I–V curves of silicon solar cells containing iron impurities can be found in [32, 66].

To create the training dataset, 25 N_{Fe} values were selected and evenly distributed on a logarithmic scale from 10^{10} cm^{-3} to 10^{14} cm^{-3}. Examples of the resulting dependencies are shown in figure 2, along with the corresponding wavelet spectrograms. The simulated test dataset consisted of 10 dependencies calculated for 10 N_{Fe} values that were not included in the training dataset.

2.3. Experiment details

The proposed method was verified using real silicon solar cells manufactured on Cz-Si:B wafers. The wafer thickness was 380 μm, and the acceptor concentration was $N_B = 1.36 \times 10^{15}$ cm^{-3}. The boron diffusion was used to shape p^+-layer (0.6 μm, 10–20 $\Omega/\square$), and the n^+-layer (0.7 μm, 20–30 $\Omega/\square$) was produced by phosphorus gas-phase mass transfer.

I–V characteristics and the I_{SC} dependencies were recorded using a Keithley 2450 source meter. Monochromatic illumination was provided by a 940 nm LED with an intensity of 5 W m^{-2}, stabilized via a W1209 thermostat in combination with a feedback-controlled power supply. The cell temperature was

Table 1. Input parameters used in the simulation of short-circuit current kinetics. $T = 340$ K.

Solar cell parameters	p^+-layer [67]	p-layer Real sample	n^+-layer [67]	
Thickness (μm)	7.75	380	0.39	
Doping concentration profile	Non-uniform	Uniform	Non-uniform	
Doping [maximum] concentration (cm^{-3})	$4.8 \cdot 10^{18}$	$1.36 \cdot 10^{15}$	$3 \cdot 10^{20}$	
Silicon properties				
Bandgap (eV)		1.1136		[68]
Bandgap narrowing* (meV)		0.84		[69]
Light absorption coefficient (cm^{-1})		227		[70]
Density of states at conduction band (cm^{-3})		2.885		[71]
Density of states at valence band (cm^{-3})		2.629		[71]
Band-to-band recombination coefficient* ($cm^3\,s^{-1}$)		$1.256 \cdot 10^{-15}$		[72]
	electron	hole		
Auger recombination coefficient* ($cm^6\,s^{-1}$)	$1.29 \cdot 10^{-30}$	$3.88 \cdot 10^{-31}$		[73, 74]
Thermal velocities ($cm\,s^{-1}$)	$2.16 \cdot 10^5$	$1.79 \cdot 10^5$		[75]
Carrier mobility* ($cm^2\,V^{-1}s^{-1}$)	1008	349		[76]
Effective mass	0.36	0.82		[77]
Defect characteristic	Fe_i	Fe_iB_s		
Level type	Donor	Donor	Acceptor	
Level energy (eV)	$E_V + 0.394$	$E_V + 0.10$	$E_C - 0.262$	[78–80]
Capture cross section electrons (cm^2)	$6.2 \cdot 10^{-15}$	$4 \cdot 10^{-13}$	$2.4 \cdot 10^{-15}$	[78–80]
Capture cross section holes (cm^2)	$8.3 \cdot 10^{-17}$	$2 \cdot 10^{-14}$	$4.4 \cdot 10^{-14}$	[78–80]
Migration energy (eV)	0.66			[40, 64, 65]
Binding energy (eV)		0.582		[63]

**Listed values correspond to the p-layer; for p^+ and n^+-layers, it was assumed that these values depend on location and are determined by the local concentration of ionized dopands.*

controlled with a thermoelectric cooler equipped with an STS-21 sensor and a PID algorithm implemented in the control software. Dissociation of FeB pairs was achieved through exposure to intense halogen lamp illumination, approximately 700 mW cm^{-2}. The illumination intervals were selected according to a previous study [81]. The kinetics of the short-circuit current were measured in the dark at 340 K for 3000 s. According to equation (3), this interval is sufficient for the complete restoration of the iron–boron pairs to their equilibrium concentration.

Figure 3(a) shows an example of the measured $I_{SC}(t)$ dependence. The signal contained some noise because, despite using a thermostat, the LED temperature fluctuated by approximately 0.4 K. A Savitzky–Golay filter was applied for smoothing, with the window lengths and filter order selected adaptively according to Krishnan and Seelamantula [52]. Only the current values corresponding to the time points used in the simulations were retained for the wavelet transformation. The smoothed curve is shown in figure 3(a), while the remaining panels of the figure display the spectrograms obtained from the raw experimental curve and the processed dependence.

To determine the iron concentration N_{Fe}, the method described in [41, 82] was employed. A total of 28 samples with iron concentrations ranging from 10^{11} cm^{-3} to 2×10^{13} cm^{-3} were examined. The entire experimental dataset was used as the test set to evaluate the models trained using the simulated

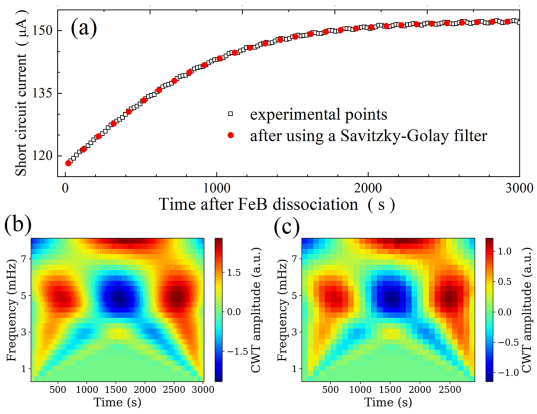

Figure 3. (a) Experimentally measured time dependence of the short-circuit current for a sample with $N_{Fe} = 2.8 \cdot 10^{13}$ cm^{-3} (a), open squares and the same dependence after applying the Savitzky–Golay filter (filled circles). Panels (b) and (c) show the wavelet spectrograms corresponding to the curves with open squares and filled circles, respectively.

data. In cases where the models were trained using experimental data, 20 randomly selected samples were included in the training set, while the remaining eight samples were used for testing.

2.4. CV models

Several CV models available in Keras were employed to extract graphical features from the wavelet spectrograms, namely EfficientNetB7, ResNet152V2, MobileNetV2, Xception, and NASNetLarge. Although these models have different architectures, they all belong to the CNN class, are designed for object classification, and have previously been successfully applied to processing EL images of solar cells [6, 43, 44, 46, 47]. Two feature extraction strategies were evaluated for all models: in the first, the class-specific probability distributions (soft labels) were passed to the subsequent stage of the pipeline, while in the second, the raw feature vectors directly extracted by the CV model were utilized.

Furthermore, the CSPDarknet53 model, which served as the CNN backbone for YOLOv4, was employed. Models of this family feature a more sophisticated CNN architecture that is not optimized for single-object classification, but for multi-object detection in images. They are widely used in imaging-based techniques [8, 42, 44]. The employed model produced three feature maps, and for subsequent processing, only the highest-level layer or the two deepest layers were selected.

It is well known that increasing feature dimensionality does not necessarily enhance the total information variance. PCA, which constructs new, uncorrelated features (principal components) was applied to mitigate the impact of redundant data. PCA is a widely used and effective technique in ML, particularly for improving performance in electrical testing techniques [83, 84]. In this study, PCA was applied to the training datasets with an explained variance threshold of 99.9%. In other words, the principal components explaining no less than 99.9% of the total variance in the original features were selected, thus achieving a substantial reduction in feature dimensionality. This pre-processing procedure was selectively applied to a subset of the CV models—specifically, those demonstrating good performance on the test sets without PCA—with the aim of assessing the feasibility and effectiveness of this approach.

Given the remarkably high dimensionality of the features produced by YOLOv4, the feasibility of applying an alternative dimensionality reduction technique was examined. Specifically, global average pooling was applied to each convolutional feature map, replacing the spatial map with its mean value, thereby yielding a single scalar value per channel.

The configurations of the CV models used in this study are summarized in table 2. The table also lists the notations that are subsequently used to refer to these configurations.

2.5. Regression algorithms

Five ML algorithms were employed to develop regression models for predicting iron concentrations: eXtreme GB (XGB), RF, SVR, GB, and DNN. The models were implemented in Python using the Keras, Scikit-learn, and XGBoost libraries.

Each regression model was trained using features obtained from all configurations listed in table 2 and subsequently used to make predictions. The only exception involved the uncompressed features extracted by YOLOv4, for which the available computational resources (2.9 GHz AMD Ryzen 7 4800H CPU, 8 GB RAM, GeForce GTX 1650 4 GB) permitted the use of SVR only. The target variable of all

Table 2. Summary of used pretrained CV models and feature extraction variants.

Base model	Model type	Feature processing	Output dimension	Model Label
EfficientNetB7	Classifier	None	1000	ENB7:CL
	Feature extractor	None	2560	ENB7:FE
		PCA	39	ENB7:FE:P
MobileNetV2	Classifier	None	1000	MNV2:CL
	Feature extractor	None	1280	MNV2:FE
		PCA	124	MNV2:FE:P
NASNetLarge	Classifier	None	1000	NAS:CL
		PCA	30	NAS:CL:P
	Feature extractor	None	4032	NAS:FE
ResNet152V2	Classifier	None	1000	R152:CL
	Feature extractor	None	2048	R152:FE
Xception	Classifier	None	1000	XCP:CL
	Feature extractor	None	2048	XCP:FE
YOLOv4 (CSPDarknet53)	Feature extractor (raw, top layer)	None	86 528	YL:FE1
		PCA	137	YL:FE1:P
	Feature extractor (raw, top & penultimate layers)	None	433 640	YL:FE2
		PCA	142	YL:FE2:P
	Feature extractor (pooled, top layer)	None	512	YL:FP1
	Feature extractor (pooled, top & penultimate layers)	None	1024	YL:FP2

models was $\log N_{\mathrm{Fe}}$. Such logarithmic transformation is a standard approach for achieving higher prediction accuracy when the target quantity spans several orders of magnitude [85, 86]. Both input features and target values were normalized to have zero mean and unit standard deviation within the training set.

Hyperparameter optimization of regression models was performed using five-fold cross-validation within the Optuna framework. The complete list of tuned hyperparameters and their respective search ranges is provided in tables S1–S5 (supplementary material). For each trial, the model was trained and evaluated across multiple folds, and the objective function was defined as the mean performance metric over all folds. In most cases, the standard deviation of the R^2 metric across cross-validation folds did not exceed 0.02, while for other metrics, the variability remained within 15% of the mean value. Collectively, these results indicate a low risk of overfitting. After hyperparameter tuning, each model was retrained on the entire training dataset using the selected hyperparameters, listed in tables S6–S10, and subsequently evaluated on a fully independent hold-out test dataset that was not involved in either tuning or calibration. Importantly, augmented versions of a given sample never appeared simultaneously in training and validation/test folds. This strategy was applied consistently to all models considered in the study.

Consequently, 87 distinct combinations of CV and regression models were investigated. Each combination was subsequently trained and evaluated using both simulated and experimental data. To identify the results for each case, a composite label was employed, derived from the last column of table 2 and the abbreviated name of the regression algorithm.

It should be noted that, due to the abstract nature of the CNN-derived feature space, the application of explainable AI techniques such as SHAP would primarily yield sensitivity measures with respect to latent dimensions rather than physically meaningful interpretations. Therefore, explainability analysis was not included in this study.

2.6. Model evaluation

A rigorous assessment of model performance across diverse metrics is essential for constructing a robust regression model. The evaluation metrics for iron quantification were the mean squared error (MSE), mean absolute percentage error (MAPE), median absolute percentage error (MedAPE) and coefficient of determination (R^2), as defined in equations (5)– (9),

$$\mathtt{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2\,, \tag{5}$$

$$\mathtt{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\mathtt{MAPE}_i\,, \tag{6}$$

$$\mathtt{MAPE}_i = \frac{|N_{\mathrm{Fe,PRED},i} - N_{\mathrm{Fe,TRUE},i}|}{N_{\mathrm{Fe,TRUE},i}} \times 100\%\,, \tag{7}$$

$$\mathtt{MedAPE} = \frac{1}{2}\left[\mathtt{MAPE}_{\lceil \frac{N}{2} \rceil} + \mathtt{MAPE}_{\lfloor \frac{N}{2}+1 \rfloor}\right], \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(N_{\mathrm{Fe,TRUE},i} - N_{\mathrm{Fe,PRED},i}\right)^2}{\sum_{i=1}^{N}\left(N_{\mathrm{Fe,TRUE},i} - \overline{N_{\mathrm{Fe,TRUE}}}\right)^2}, \tag{9}$$

where $\hat{y_i}$ is the predicted value of the target variable for the ith data point, y_i is the corresponding known value (obtained by logarithmic transformation and normalization of the iron concentration); N denotes the number of samples in the dataset ($N = 25$ for the simulated training set, 20 for the experimental training set, 10 for the simulated test set, and 28 and 8 for the experimental datasets used to test models trained on the simulated and experimental sets, respectively); $\mathtt{MAPE}_i$ is the absolute percentage error for ith data point; $N_{\mathrm{Fe,PRED},i}$ is the predicted iron concentration, $N_{\mathrm{Fe,TRUE},i}$ is the known value (either the parameter used in the simulation or obtained from experimental iron determination); equation (8) implies that $\mathtt{MAPE}_i$ must be arranged in order of magnitude; $\overline{N_{\mathrm{Fe,TRUE}}}$ is the mean of the true values in the dataset.

MSE is one of the most widely used metrics for evaluating model accuracy, and the training objective was specifically defined to minimize this quantity. However, because the computation of y_i involves both normalization and logarithmic transformation of $N_{\mathtt{Fe}}$, the MSE metric alone does not fully reflect the accuracy of the iron concentration estimation. Therefore, MAPE, which quantifies the mean relative deviation, was also employed. In addition, MedAPE, representing the error value below which half of the predictions lie, provides a more robust measure against the influence of individual outliers, which can have a particularly large effect on mean-based metrics in smaller datasets. Finally, the R^2 was used to quantify the fraction of the variance in the target variable explained by the model, thereby indicating how well the predicted values reproduce the observed data; a value of 1 corresponds to perfect agreement.

Confidence intervals were estimated using a hierarchical resampling procedure. Regression models were trained in five independent runs with different random initializations and stochastic optimization trajectories while maintaining a fixed set of hyperparameters determined during the tuning phase. For each trained model, performance was evaluated on 100 bootstrap resamples of the test dataset generated with replacement. Confidence intervals were then derived from the empirical distribution of the resulting performance metrics. The absence of reported confidence intervals for certain model variants indicates that their magnitudes fall below the reporting precision of the displayed results.

3. Results and discussion

3.1. Simulated data

Figure 4 illustrates the representative prediction results obtained from the models trained using the simulated training dataset. The complete set of results covering all the 87 investigated configurations is provided in figure S1 of the supplementary material.

Supplementary figures S1–S10 present a comprehensive set of performance metrics and predicted-versus-true dependencies for all 87 considered combinations of CV feature extractors and regression models. While the main text shows representative examples, the supplementary materials confirm that the observed trends and relative model rankings remain consistent across all evaluated scenarios.

Figure 5 presents the MAPE and R^2 scores obtained when the models were applied to the training dataset. Although these metrics are sufficiently representative, the corresponding MedAPE and MSE values are shown in figure S2 (supplementary material). The most notable observation from figures 4 and 5 is that, despite the very limited size of the training dataset (25 samples), most models demonstrate high training performance; in many cases, the mean relative error is approximately 1% or lower, and it rarely exceeds 10%. At the same time, the R^2 score drops below 0.980 in only 8 out of 87 cases. It is also worth noting that the MedAPE values generally do not exceed the MAPE and are, in fact, smaller in

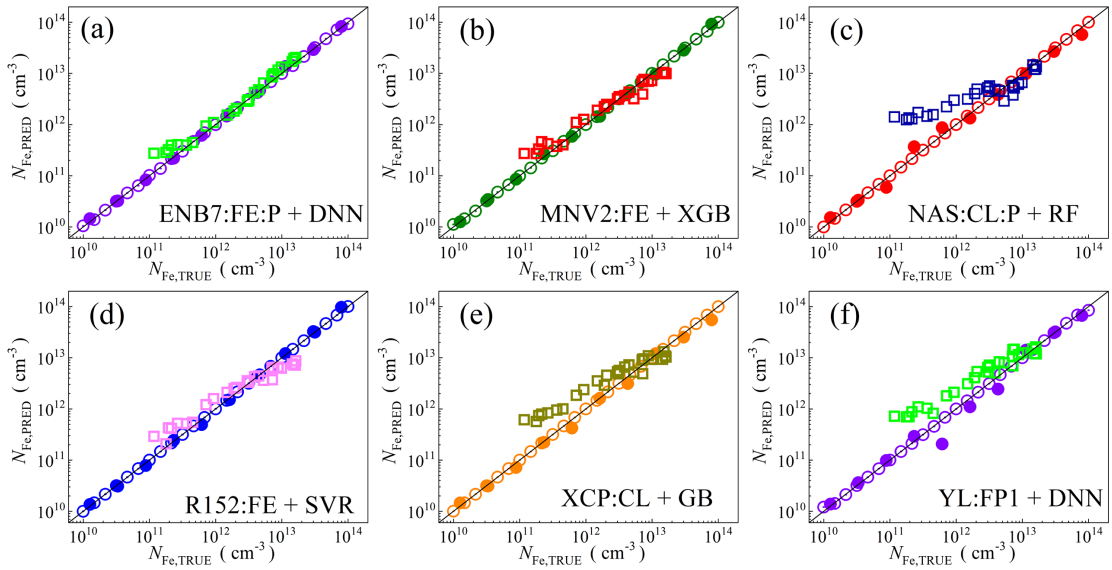

Figure 4. Scatter plots compare the reference iron $N_{Fe,TRUE}$ with ML-predicted values $N_{Fe,PRED}$, obtained using feature vectors extracted from various CV models combined with different regression algorithms (specific models are indicated in the figures). ML models were trained using a simulated dataset. The open circles correspond to the training phase, whereas the filled circles and open squares correspond to the test phase representing the simulated and experimental datasets, respectively. The black lines indicate the identity line between predicted and true values.

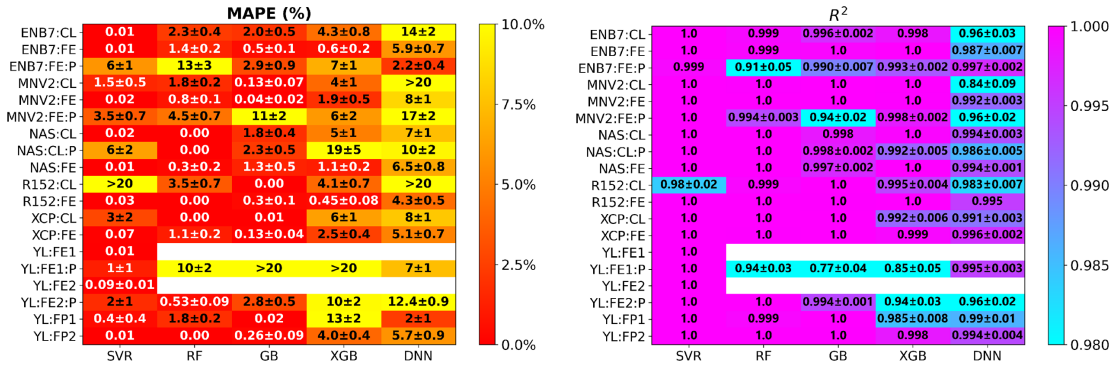

Figure 5. Mean absolute percentage error (left panel) and coefficient of determination (right panel) for different combinations of CV models (vertical axis) and regression models (horizontal axis) during the training phase. The models were trained using the simulated dataset.

most instances (see figure S2). These consistently high performance metrics indicate that (i) the wavelet transformation produced highly informative images that effectively encoded information about the Fe concentration, and (ii) the CV models successfully extracted features correlated with the concentration.

Among the regression models, GB and SVR exhibited the best performance, whereas DNN produced the least favorable results. This outcome is entirely consistent with expectations because both GB and SVR are well known to perform effectively when the number of samples is limited and the feature space exhibits a low noise level, as is characteristic of synthetic datasets. In contrast, neural networks contain a large number of parameters and therefore do not tend to exhibit perfect generalization under such conditions.

Among the CV models, EfficientNetB7 and NASNetLarge demonstrated the best performance, whereas ResNet152V2 and YOLOv4 exhibited the weakest results. This discrepancy can be attributed to the fact that the former two are relatively modern architectures specifically designed to extract generalizable features, and are therefore well suited to wavelet spectrograms, which are characterized by a multi-scale structure. In contrast, ResNet152V2 is primarily optimized for object classification, whereas YOLO is less effective for regression tasks involving global image patterns, as it focuses on localized object detection.

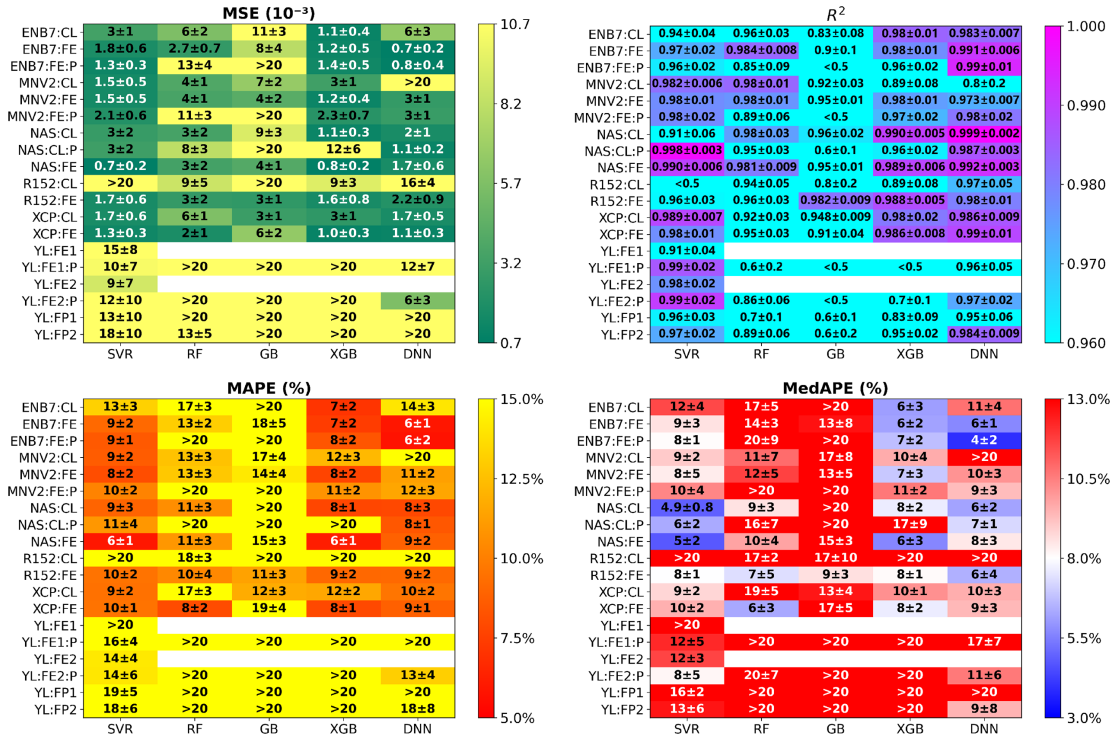

Figure 6. Mean squared error, coefficient of determination, mean absolute percentage error, and median absolute percentage error for different combinations of CV models (vertical axis) and regression models (horizontal axis) during the test phase with the simulated dataset. The models were trained using a simulated training dataset.

It is also evident that using class probabilities as descriptors degrades the prediction quality compared with the cases in which image features are employed directly. This indicates that the internal feature maps of the CNN models provide a more informative representation of visual patterns, enabling the regressor to establish a stronger relationship with concentration. Moreover, the application of PCA, even though it retains 99.9% of the variance, results in reduced predictive accuracy. This observation suggests that the image patterns associated with variations in iron concentration may constitute only a minor portion of the overall data variance.

The ability of models to achieve high accuracy on the training set is a prerequisite for effective performance on unseen data; however, this does not guarantee reliable prediction outcomes. Consequently, the evaluation of an independent test set is essential. This step becomes particularly critical when the training set is small because the test results provide primary evidence for model generalizability. Figures 4 and 6 present the prediction results obtained for the test set generated from synthetic data (detailed versions are available in figures S1 and S3 of the supplementary material). As expected, prediction performance declined. Nevertheless, we emphasize the high R^2 values and the small gap between the training and test R^2 values (less than 0.05). These results clearly indicate that the models did not rely on memorization but instead successfully captured meaningful underlying patterns in the data.

The reduction in accuracy varied across the regression algorithms. Specifically, this degradation was least pronounced for the DNN, which achieved the best overall performance. The poorest metrics were obtained for RF and GB, whereas SVR and XGB performed slightly worse than DNN but with a relatively small margin. This behavior can be attributed to the ability of the DNN to smoothly approximate the continuous dependencies. In contrast to GB and RF, which rely on the formation of local decision rules, neural networks construct a continuous surface in the feature space. This property enhances interpolation for iron concentration values not represented in the training data. From this perspective, XGB and SVR occupy an intermediate position, explaining their slightly lower performance relative to the DNN.

The relative performance of the CV models on the simulated test dataset remained consistent, with no change in the top- and bottom-performing architectures. Specifically, EfficientNetB7 and NASNetLarge yielded the most favorable results, confirming their effectiveness in extracting wavelet image features relevant to concentration prediction. By contrast, ResNet152V2 and YOLOv4 produced

Table 3. Summary of 1D-CNN model performance metrics.

Training data type	Evaluation dataset	MSE (10^{-3})	R^2	MAPE (%)	MedAPE (%)	Reference (CV-based Models)[a]
Simulated	Train	31 ± 5	0.89 ± 0.02	35 ± 5	32 ± 2	Figure 5 and S2
	Test	30 ± 5	0.90 ± 0.04	35 ± 5	32 ± 3	Figure 6 and S3
	Experimental, without post-hoc calibration	530 ± 70	< 0.1	480 ± 60	400 ± 20	Figures 8(a), (b) and S4
	Experimental, with post-hoc calibration	130 ± 30	0.6 ± 0.1	50 ± 10	44 ± 8	Figures 8(c), (d) and S5
Experimental	Train	27 ± 7	0.2 ± 0.1	33 ± 6	26 ± 8	Figure 10 and S7
	Test	30 ± 10	< 0.1	40 ± 10	30 ± 15	figure 11 and S8

[a] The column indicates the figures in which performance metrics obtained for CV-based models under comparable conditions are reported.

the least satisfactory outcomes. A notable exception arises in configurations where YOLOv4 features from two layers are combined with either a DNN or SVR. In these cases, performance metrics were considerably improved. This enhancement suggests that the use of a larger number of features enabled the capture of more diverse patterns, which, when coupled with flexible regressors, partially mitigated the limitations observed in the standalone YOLO model.

Interestingly, the PCA application to the test set did not substantially degrade performance; in certain cases, it even produced slight improvements. This indicates that PCA is not universally detrimental, although the resulting gains are generally marginal and unpredictable. Moreover, the differences in metrics between models using class probabilities and those employing raw image features were less pronounced than observed for the training dataset. Nevertheless, uncompressed feature representations consistently maintained a performance advantage.

The top-performing model combinations were identified as follows: ENB7:FE+DNN (MAPE and MedAPE 6%, $R^2 = 0.99$), ENB7:FE:P+DNN (MAPE = 6%, MedAPE = 7%, $R^2 = 0.99$), NAS:FE+SVR (MAPE = 6%, MedAPE = 5%, $R^2 = 0.99$), NAS:FE+XGB (MAPE = 6%, MedAPE = 5%, $R^2 = 0.99$), and NAS:CL:P+SVR (MAPE = 11%, MedAPE = 5%, $R^2 = 0.99$). These results correspond to exceptionally high absolute performance metrics, providing clear evidence that the models effectively capture the underlying relationship between the wavelet-based images and the iron concentration.

To verify the feasibility of using CV-based models, CNN architectures were also designed to directly process the kinetic dependencies of the short-circuit current, without wavelet transformation, in order to predict iron concentration (hereinafter referred to as the 1D-CNN). The structure of the model was identical to that previously used to successfully analyze one-dimensional signal associated with defects [19]. Specifically, the 1D-CNN model consists of two one-dimensional convolutional layers, followed by global average pooling and three fully connected layers. The convolutional layers use a kernel size of three and padding of one, with eight and sixteen output channels, respectively. The pooled features are processed by fully connected layers with 64 and 32 neurons, and the final prediction is produced by a single-neuron output layer. During model tuning, batch normalization and regularization in the convolutional layers, dropout between fully connected layers, activation functions, and weight initialization were optimized. The models were trained and tested on the same dataset, comprising both artificial and experimental data, from which the wavelet spectrograms were generated. The resulting performance metrics are presented in table 3, where the first two rows correspond to the cases discussed above for models that include both regression and CV components.

First, a small difference between the metrics obtained for the training and test sets is observed. This behavior suggests an appropriate model complexity of the 1D-CNN, similar distributions of the training and test data, and a correct train-test split. At the same time, the absolute values of the metrics are low, indicating underfitting, which is expected given the extremely small size of the training dataset. Moreover, these values are substantially lower than those achieved by the best CV based models. This result suggests, on the one hand, that CV based transfer learning is highly effective and, on the other hand, that not every CV model is suitable for identifying features relevant to the restructuring of iron-containing defects.

It should be noted that solar cells typically operate over a broad temperature range. Temperature influences the characteristic time of FeB complex association (see equation (3)), the thermodynamic equilibrium between Fe_i and FeB defect concentrations, the carrier capture cross-sections of these

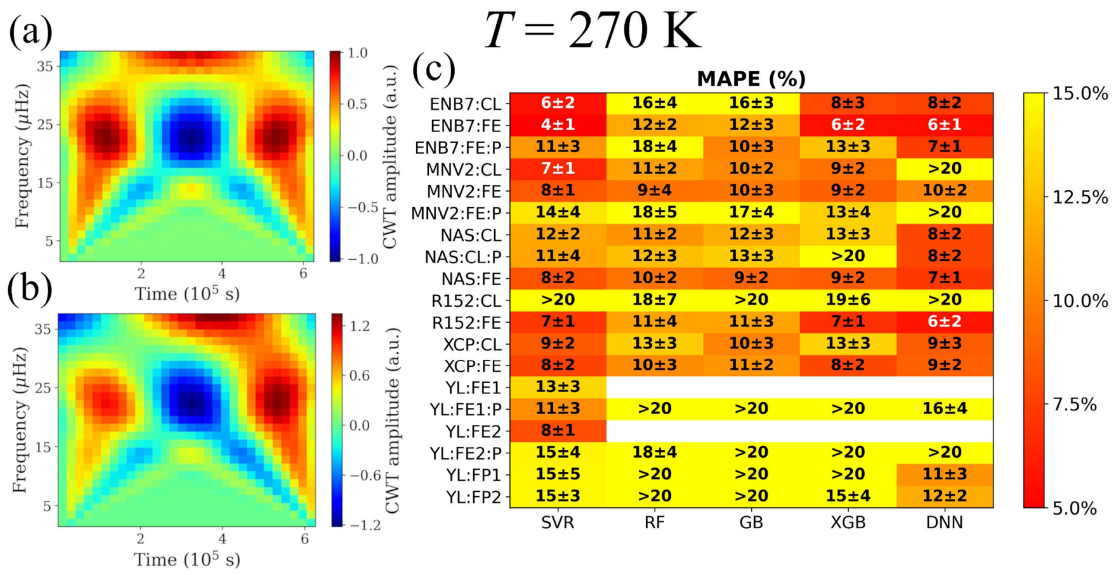

Figure 7. Wavelet spectrograms of simulated short-circuit current time dependencies at 270 K for iron concentrations of 10^{10} cm^{-3} (a) and 10^{14} cm^{-3} (b). Mean absolute percentage error (c) for different combinations of CV models (vertical axis) and regression models (horizontal axis) evaluated during the test phase using the dataset simulated at 270 K.

defects, which together determine the recombination activity of iron-related centers, as well as the silicon parameters that govern the photogeneration and transport of excess carriers. Collectively, these factors affect the kinetics of the short-circuit current and its dependence on the iron impurity concentration. This raises the question of whether temperature may also impact the performance of ML models.

Simulations were performed for 270 K. As an example, figures 7(a) and (b) show the wavelet spectrograms for two distinct iron concentrations. Comparison with figures 2(b) and (c) indicates that, at lower temperatures, changes in iron concentration produce more pronounced differences in the spectrograms, suggesting a corresponding increase in model prediction accuracy. Figure 7(c) presents the MAPE for models trained and tested on synthetic datasets prepared in the same manner as those used previously, but under 270 K conditions. More detailed metrics are provided in figures S4 and S5 of the supplementary materials. Comparison with figure 6 (lower right panel) confirms that a notable improvement in model performance is observed.

However, it should be emphasized that the proposed methodology for iron concentration assessment is primarily intended for practical application. Calculations based on equation (3) indicate that τ_{ass} is approximately 850 s at 340 K, whereas it increases to about 230 000 s at 270 K. The simulated $I_{SC}(t)$ dependencies at 340 K assume that the entire curve can be measured within 50 min. At 270 K, however, the interval between consecutive points extends to 21 600 s, corresponding to six measurements per day and requiring nearly one working week to acquire a complete curve. Even at 300 K, the characteristic time of 15 000 s remains prohibitively long. Consequently, although lower measurement temperatures could enhance prediction accuracy, experimental validation under such conditions is impractical. Therefore, hereafter this study focuses exclusively on results obtained at 340 K, both computationally and experimentally.

Having established the effectiveness of the models on the simulated test dataset, we evaluated their performances using experimental measurements. This step is critical for assessing the generalizability of the models to real-world data where additional sources of noise and variability may be present. By comparing the results obtained from the experimental data with those from the simulated dataset, it is possible to identify the potential limitations of the models and to confirm whether the features extracted from wavelet spectrograms remain informative under practical conditions. Figures 8(a) and (b) show the performance metrics of models trained on synthetic data when applied to experimental measurements. The relationship between predicted and actual concentrations is illustrated in figure 4, with additional results provided in figures S1 and S6. As observed, the mean and median prediction errors fall within the (15–25)% range for only a limited number of configurations, specifically certain combinations of EfficientNetB7 or NASNetLarge with DNN or SVR. Although this outcome is not catastrophic, considering the approximately 10% inherent experimental error in N_{Fe} determination, it falls short of ideal

expectations. At the same time, the R^2 metric remains acceptably high. Further analysis (figure 4) indicates that the prediction error depends on the iron concentration level: the relationship between $N_{\mathrm{Fe,PRED}}$ and $N_{\mathrm{Fe,TRUE}}$ is linear on a logarithmic scale, but its slope deviates from the line of unity. It is essential to emphasize that the synthetic data were generated using explicit and transparent physical laws, and that the calculations employed realistic parameter values reported in the literature. Taken together, these considerations indicate that the observed discrepancy represents a systematic prediction bias rather than a loss of correlation or a failure to capture the underlying physical relationships, implying that the models are capable of capturing relative differences in concentration but do not accurately predict absolute values.

Although the presence of residual noise patterns in the experimental curves, which were insufficiently suppressed by filtering, could contribute to the observed discrepancies, a more plausible explanation lies in the incomplete correspondence between the physical model used for data synthesis and the actual behavior of the solar cells. This mismatch is likely associated with the numerical parameters employed in the fundamental equations of the model (equations (1)– (3)). For instance, the calculation of the characteristic FeB association time (equation (3)) adopted values of $A = 5.7 \times 10^5 \frac{\mathrm{s}}{\mathrm{K\,cm^3}}$ and $E_{\mathrm{m}} = 0.66$ eV, which are among the most frequently reported in the literature. However, a considerable scatter exists in these parameters. Specifically, E_{m} values ranging from 0.55 eV [87] to 0.69 eV [88] have been reported, including intermediate estimates of 0.64 eV [89], 0.65 eV [39], 0.66 eV [64, 65, 90, 91], 0.67 eV [92], and 0.68 eV [63, 93, 94]. Similarly, the pre-exponential factor A has been cited as 4.3×10^5 [95] or 5×10^5 [91, 96]. Moreover, the calculations assumed a spatially invariant E_{m} across the device. In practice, the reported diffusion barriers are typically derived for bulk p-Si, whereas the energy value can vary in regions with different Fermi level positions [97], such as the space-charge region in the present structures. A similar variability exists for the other key parameters. The FeB pair binding energy has been reported to range from 0.45 to 0.67 eV [39, 63, 92, 98], the Fe_i donor level position varies between 0.38 and 0.394 eV above the valence band maximum [40, 93, 99, 100], and the pre-exponential factor in denominator of equation (2) A_z has been cited as either 10^{-23} cm^3 or 2.7×10^{-22} cm^3 [92]. A deviation of any of these parameter values from their actual physical magnitudes could account for the observed prediction errors. Furthermore, earlier studies have demonstrated that both the non-uniform distribution of iron across the base thickness and the variation in the base thickness itself can significantly affect iron concentration estimation [39]. Neither of these effects was incorporated into the present modeling framework.

A common strategy for improving prediction accuracy involves post-hoc calibration, in which a corrective function is applied to model outputs using parameters derived from a limited subset of experimental data. In the present case, the analysis indicated that quadratic correction of the target variable provided the most suitable adjustment, expressed as follows:

$$\log N_{\mathrm{Fe,PRED}} = 9.51 - 1.71 \cdot \log N^{*}_{\mathrm{Fe,PRED}} + 0.079 \cdot \left(\log N^{*}_{\mathrm{Fe,PRED}}\right)^2, \tag{10}$$

where $N^{*}_{\mathrm{Fe,PRED}}$ denotes the direct model prediction. The performance metrics after post hoc calibration are presented in figures 8(c) and (d). As shown by the data, applying this correction substantially reduced the prediction errors. Specifically, the mean relative error across the 28-sample experimental dataset now lies within the (13–17)% for the best-performing models (approximately 20 out of 87 configurations) and remains below 25% for most of the others. The median error is even lower, reaching (8–10)% in the most favorable cases (figure S7). In line with the results obtained for the simulated test dataset, the most accurate configurations are EfficientNetB7, NASNetLarge, DNN, SVR, and XGB. Interestingly, MobileNetV2 was also among the top performing combinations. This outcome may suggest that the features extracted by MobileNetV2, although less informative for the training and simulated test datasets, capture specific image patterns more relevant to the experimental measurements and applied correction could have further enhanced the effectiveness of these features.

It is worth noting that applying the correction increased the gap between the mean and MedAPEs. This observation suggests that, although the correction improved the overall agreement between the predicted and true values, a few samples still exhibited relatively large residual errors.

This correction also led to a decrease in the R^2 value. This outcome is expected because post-processing can weaken the linear correspondence between the initial predictions and experimental values, even when the overall prediction errors are simultaneously reduced. Therefore, the reduction in R^2 should be regarded as a side effect of enhancing the accuracy of the model, rather than as evidence of its degradation.

The third and fourth rows of table 3 present the performance metrics obtained by the 1D-CNN when a model trained on synthetic data was applied to experimental data. The initial results obtained

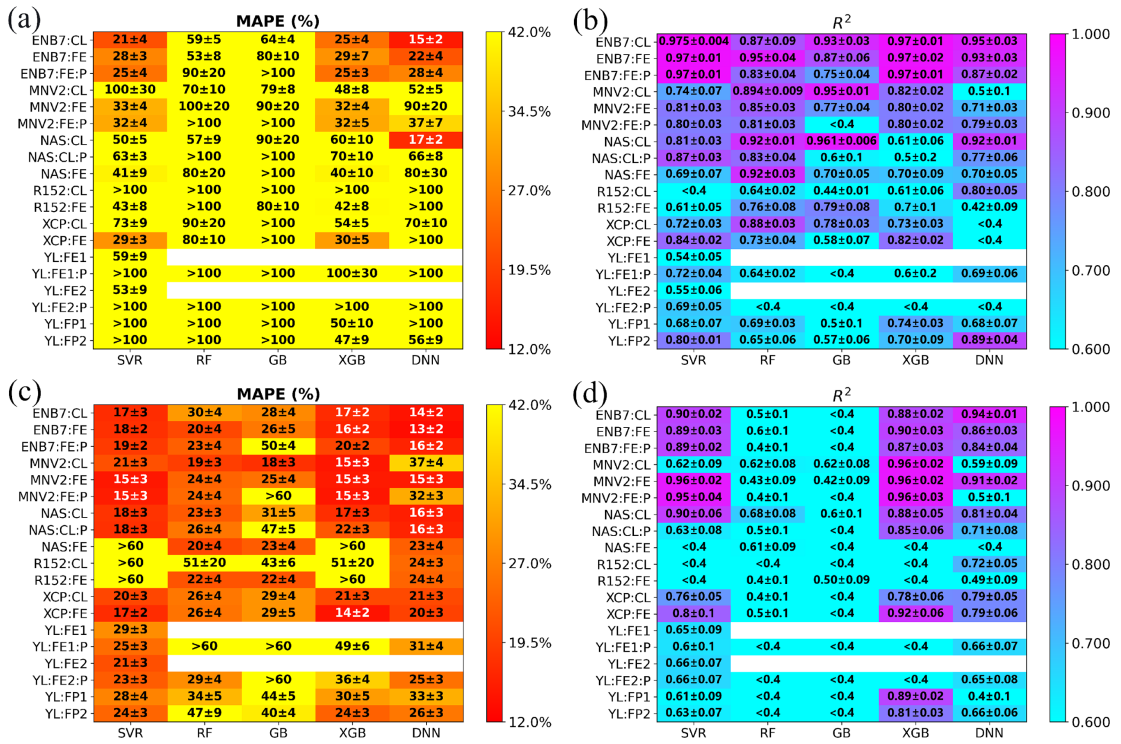

Figure 8. Mean absolute percentage error (a), (c) and coefficient of determination (b), (d) for different combinations of CV models (vertical axis) and regression models (horizontal axis) during the test phase with the experimental dataset without (a), (b) and with (c), (d) post hoc calibration according to equation (10). The models were trained using a simulated dataset.

without post-hoc calibration are notably poor, which clearly demonstrates the need to incorporate transfer learning techniques. Notably, the post-hoc calibration, specifically the function defined by equation (10) and derived from the CV based model results, leads to a substantial improvement in the predictive performance of the 1D-CNN. This observation points to two key conclusions. First, it reveals the presence of a systematic bias between the raw model outputs and the true values, which appears to be independent of the specific model architecture. Second, it indicates that the dominant source of error does not originate from the internal structure of the models, which are capable of learning meaningful patterns and reproducing structural dependencies, but rather from the properties of the training data, in particular the mismatch between synthetic and experimental datasets. This finding provides a dual perspective. On the one hand, it opens a pathway for refining defect parameters by adjusting the values used in simulations until post-hoc calibration is no longer required. On the other hand, it strongly suggests that the commonly accepted parameter values for iron-containing defects employed in the simulations may be intrinsically inaccurate.

In section summary, SCAPS-1D simulations may not reproduce all quantitative details of the short-circuit current kinetics due to deviations from precise experimental conditions. The primary objective of using the artificial dataset was to demonstrate that CV models can extract physically meaningful features from wavelet-transformed representations, enabling the estimation of uncontrolled metallic impurity concentrations. This goal was successfully achieved using several representative CV models, including EfficientNetB7 and NASNetLarge. The post-hoc calibration procedure equation (10) is purely heuristic, improving predictive accuracy primarily within the training range but providing no physical insight into the underlying system. Its applicability beyond the calibrated domain is therefore limited, and it does not correct structural model errors. A more robust approach, implemented later in this study, integrates experimental data directly into the training process, enhancing model transferability to real measurements.

3.2. Experimental data

Figure 9 shows the correlation between the iron concentrations predicted by the models trained on the experimental data and the reference values derived according to [41, 82]. Representative outcomes for selected CV-regression model combinations are shown for both training and testing, with the extended data provided in figure S8 of the supplementary material.

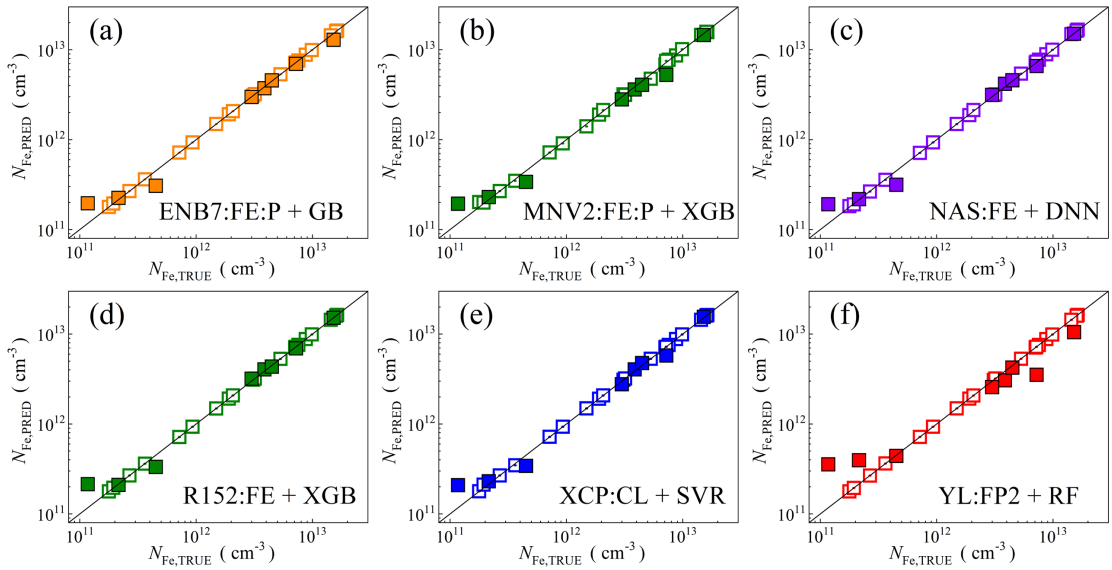

Figure 9. Scatter plots compare the reference iron concentrations $N_{Fe,TRUE}$ with ML-predicted values $N_{Fe,PRED}$, obtained using feature vectors extracted from various CV models combined with different regression algorithms (specific models are indicated in the figures). ML models were trained using a dataset derived from experimental measurements. The open and filled squares correspond to the training and testing phases, respectively. The black lines indicate the identity line between predicted and true values.

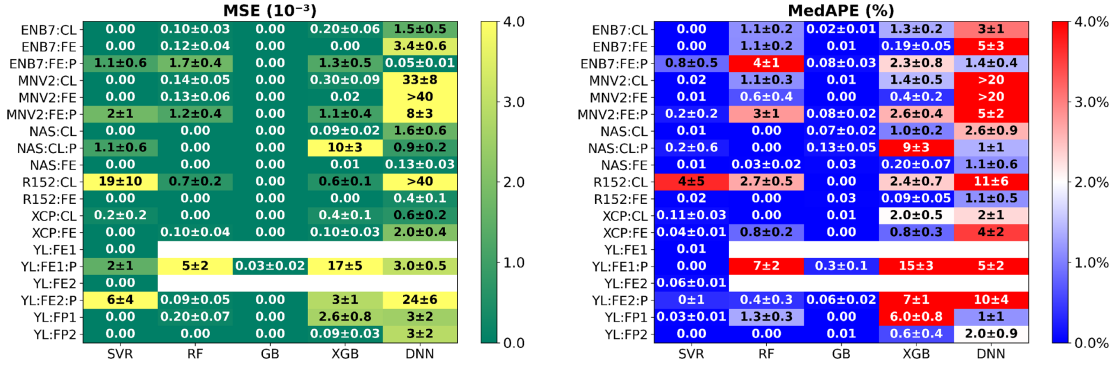

Figure 10. Mean squared error (left panel) and median absolute percentage error (right panel) for different combinations of CV models (vertical axis) and regression models (horizontal axis) during the training phase. The models were trained using an experimental dataset.

It should be noted that, compared to training on synthetic data, this case employed an even smaller sample set (20 samples versus 25). However, these samples covered a narrower range of iron concentrations (10^{11}–$2 \cdot 10^{13}$ cm^{-3} versus 10^{10}–10^{14} cm^{-3}). Figure 10 presents a subset of the performance metrics obtained during the training phase (a more comprehensive overview is provided in figure S9). Overall, the results were similar to those shown in figure 5. In many cases, extremely low errors (below 0.5%) and high coefficients of determination (approaching unity) were observed. This behavior is particularly characteristic of GB, RF, and SVR. As in the previous case, the DNN exhibited weaker performance relative to the other algorithms, although its results slightly improved compared to those for the simulated training dataset. This improvement may stem from the greater homogeneity of the experimental data. The leading performance of EfficientNetB7 and NASNetLarge among the CV models was again confirmed, indicating that these architectures produce the most relevant features for this task. However, the performance gap relative to other CV models has become less pronounced.

Interestingly, the application of PCA often enhances the performance of the DNN. For example, the MAPE values for ENB7:FE and ENB7:FE:P were 7% and 1%, respectively. This observation suggests that PCA effectively mitigated the adverse effects of high feature dimensionality when the sample size was limited. In contrast, no similar improvements were observed for other regression algorithms. Moreover, for the experimental training dataset, the difference between MAPE and MedAPE becomes smaller. This reduction implies a more symmetric error distribution and a decreased occurrence of extreme deviations.

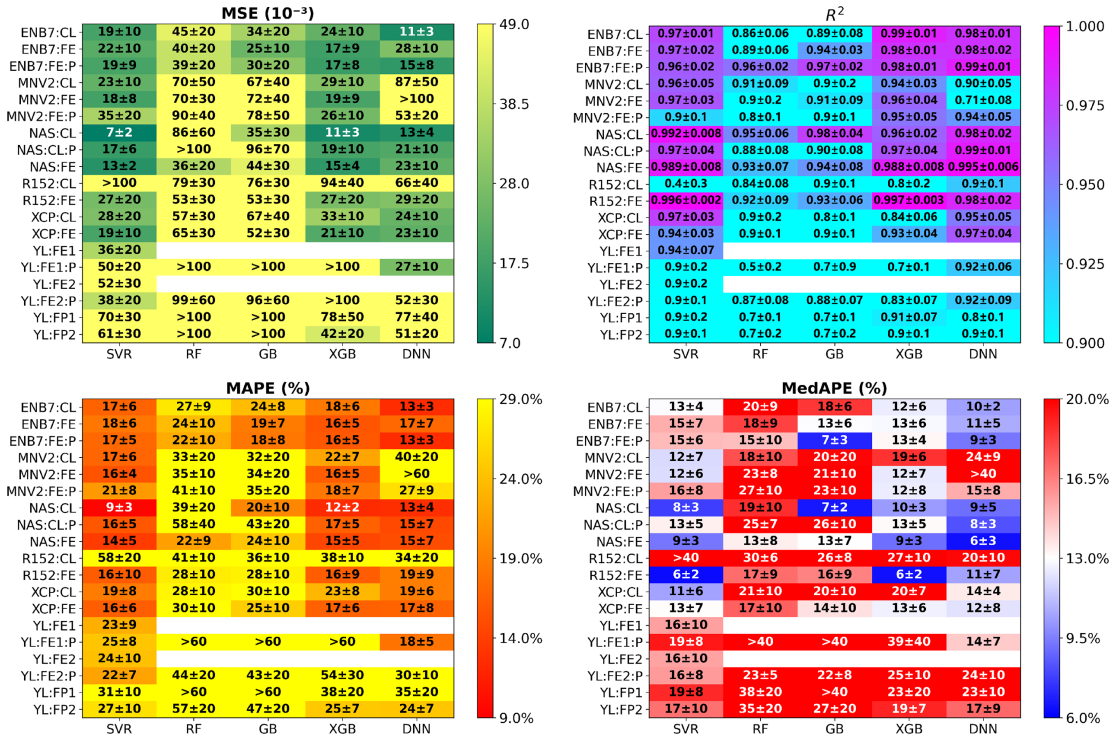

Figure 11. Mean squared error, coefficient of determination, mean absolute percentage error, and median absolute percentage error for different combinations of CV models (vertical axis) and regression models (horizontal axis) during the test phase. The models were trained using an experimental dataset.

Figures 11 and S10 show heatmaps of the prediction metrics for the test experimental dataset using models trained on a separate subset of experimental data. Compared with the models trained on simulated data, the predictive performance of the experimental test dataset improved substantially. In particular, for several of the best-performing CV-regressor combinations, the MAPE and MedAPE values fall within (6–15)%, which is markedly better than the 20%–30% observed previously. Moreover, the R^2 values predominantly exceed 0.97, indicating a strong agreement with the actual dependencies. The accuracy achieved through direct training on experimental data is comparable to that obtained with post-hoc correction; however, the higher correlation coefficients suggest a more faithful representation of both the scale and the variations of the underlying dependency. SVR, DNN, and XGBoost remain the top-performing regressors, whereas EfficientNetB7, NASNetLarge, and—surprisingly—ResNet152V2 (when using image features) are the strongest CV models. YOLOv4 and MobileNetV2 consistently exhibit the weakest performance. Interestingly, for experimental data, using class probabilities as descriptors (:CL) yields results comparable to, and in some cases slightly better than, direct image features (:FE), particularly when strong CV architectures are combined with flexible regressors. This likely reflects the more compact and aggregated nature of class features, which makes them less sensitive to experimental noise. Conversely, for weaker CV models, the :FE configurations retain their advantage, consistent with previous findings. Finally, applying PCA to :FE models improved prediction accuracy, whereas it did not benefit models based on class features. This suggests that reducing the dimensionality of highly noisy features contributes to improving prediction accuracy on experimental data.

The performance metrics of the 1D-CNN trained exclusively on experimental data are presented in the last two rows of table 3. These metrics are clearly inferior to those obtained by most models that incorporate pre-trained CV components. This observation further underscores the effectiveness of the proposed transfers learning methodology for applications involving limited datasets.

In summary, three approaches for predicting iron concentrations from experimental data were evaluated: (i) models trained on simulated data, (ii) models trained on simulated data with post hoc correction, and (iii) models trained directly on experimental data. Training on simulated data alone yielded moderate agreement with experimental measurements; however, systematic biases were observed owing to differences between the synthetic and real systems. Post-hoc correction effectively reduced these biases, lowering the mean and median errors, yet the correlation with the actual variation remained limited.

Table 4. Overview of supplementary figures S1–S10*.

Figures	Training dataset	Evaluation dataset/condition	Content	Key message
S1, S8	Simulated (S1), experimental (S8)	Simulated (S1), experimental (S1, S8)/training & test	Correlation between $N_{Fe,TRUE}$ and $N_{Fe,PRED}$ (scatter plots)	Identification of systematic biases and prediction linearity
S2–S3	Simulated	Simulated/training (S2), test (S3)	Heatmaps of MSE, R^2, MAPE, and MedAPE	Baseline comparison and ranking of all models
S4–S5	Simulated ($T = 270$ K)	Simulated ($T = 270$ K)/training (S4), test (S5)	Heatmaps of performance metrics	Impact of specific temperature conditions (270 K) on model accuracy
S6–S7	Simulated	Experimental/test without (S6) or with (S7) post-hoc calibration	Performance heatmaps	Effect of post-hoc calibration on experimental predictions
S9–S10	Experimental	Experimental/training (S9), test (S10)	Accuracy heatmaps	Validation of the purely empirical training pipeline

*All figures report results for all 87 combinations of CV feature extractors and regression models.

Direct training on experimental data provided the most balanced outcome, achieving both low prediction errors and high correlation coefficients, thereby demonstrating the importance of incorporating real measurements during model development. Thus, training directly on experimental data improves prediction accuracy and eliminates the systematic biases observed in models transferred from synthetic data. At the same time, achieving optimal performance requires both the careful selection of the CV model-regressor combination and the inclusion of real experimental data.

However, it should be noted that scaling the short-circuit current at every point of the kinetic curve by a constant factor does not affect the graphical representation of the resulting wavelet spectrogram; only the amplitude of the CWT is modified, but this change is proportional across all frequency and time values. Consequently, such transformation does not influence the features extracted by the CV model. In practice, such a transformation may result from the use of a linear signal amplifier or from the presence of shunt and/or series resistances. Therefore, the proposed approach is inherently resistant to parasitic resistances, which represents an additional advantage.

4. Conclusion

This study demonstrates that transfer learning from pretrained CV models enables accurate regression modeling, even with extremely small training datasets typical of experimental materials research. The proposed workflow involves measuring a characteristic kinetic dependency, transforming it into an image via wavelet analysis, extracting features using a pretrained CV model, and training a regression model on these features to predict material properties. The feasibility of this approach is illustrated by predicting the iron impurity concentration in silicon solar cell from short-circuit current kinetics following FeB pair dissociation using a training dataset of only 20–25 samples.

The performances of models trained on both synthetic and experimental datasets were evaluated. In both cases, EfficientNetB7 and NASNetLarge provided the most informative features, whereas DNNs and SVR yielded the highest prediction accuracy. The best-performing models achieved MSE, MAPE, MedAPE, and R^2 values of 0.001, 6%, 4%, and 0.999, respectively, for synthetic data, and 0.008, 10%, 5%, and 0.996, for experimental data.

When training models on synthetic data, image feature vectors serve as the most suitable descriptors for the regression model. In contrast, when experimental data are used, prediction accuracy can be improved and the influence of noise reduced by utilizing class probability distributions or applying PCA.

The combination of transfer learning from CNNs with an appropriate choice of descriptor type and regression algorithm represents a promising strategy for materials research in general and for defect characterization in particular, especially when the acquisition of large datasets is difficult or impractical.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/olegolikh/CV_Fe_SiSC.git [101].

Supplementary data 1 available at https://doi.org/10.1088/1361-6641/ae3850/data1.

Conflict of interest

The authors declare that there are no conflicts of interest related to this work.

Ethics statement

This study does not involve human participants, animals, or procedures requiring ethical approval.

Author contributions

Oleg Olikh 0000-0003-0633-5429
Conceptualization (equal), Formal analysis (equal), Investigation (equal), Methodology (equal), Software (equal), Supervision (equal), Visualization (equal), Writing – original draft (equal), Writing – review & editing (equal)

Oleksii Zavhorodnii 0000-0001-8080-7661
Investigation (equal), Software (equal), Visualization (equal), Writing – review & editing (equal)

Yulia Perets 0000-0003-0567-770X
Data curation (equal), Project administration (equal), Software (equal), Validation (equal)

References

[1] International Energy Agency 2024 Renewables 2024 *Technical Report* (IEA Paris) (licence: CC BY 4.0) (available at: www.iea.org/reports/renewables-2024)
[2] Osama A, Tina G M, Gagliano A, Jimenez-Castillo G and Munoz-Rodríguez F J 2025 *Sol. Energ. Mat. Sol.* **287** 113625
[3] Institute P 2025 Solar generation grew by 30% in 2024, says IEA (available at: https://prometheus.org/2025/02/23/solar-generation-grew-by-30-in-2024-says-iea/ (accessed 10 October 2025)
[4] Fischer M, Woodhouse M, Brammer T and Puzant B 2025 International technology roadmap for photovoltaic (ITRPV) *Technical Report* (VDMA e. V Frankfurt am Main)
[5] Thome F T, Garashli E, Kwapil W, Schindler F and Schubert M C 2025 *Sol. Energy Mater. Sol. Cells* **293** 113854
[6] Jia Y, Chen G and Zhao L 2024 *Sci. Rep.* **14** 15170
[7] Hijjawi U, Lakshminarayana S, Xu T, Piero Malfense Fierro G and Rahman M 2023 *Sol. Energy* **266** 112186
[8] Liu Y, Wu Y, Yuan Y and Zhao L 2024 *Opt. Express* **32** 17295–317
[9] Doll B, Hepp J, Hoffmann M, Schuler R, Buerhop-Lutz C, Peters I M, Hauch J A, Maier A and Brabec C J 2021 *IEEE J. Photovolt.* **11** 1419–29
[10] Datta S, Baul A, Sarker G C, Sadhu P K and Hodges D R 2023 *IEEE Access* **11** 77750–78
[11] Jaiswal R, Martinez-Ramon M and Busani T 2023 *IEEE J. Photovolt.* **13** 2–15
[12] Buratti Y, Javier G M N, Abdullah-Vetter Z, Dwivedi P and Hameiri Z 2024 *Renew. Sust. Energy Rev.* **202** 114617
[13] Mahdavipour Z 2024 *Sol. Energ. Mat. Sol.* **278** 113210
[14] Hopwood M W, Gunda T, Seigneur H and Walters J 2020 *IEEE Access* **8** 161480–7
[15] Li B, Delpha C, Diallo D and Migan-Dubois A 2021 *Renew. Sust. Energy Rev.* **138** 110512
[16] Liu Y, Ding K, Zhang J, Li Y, Yang Z, Zheng W and Chen X 2021 *Energy Convers. Manage.* **245** 114603
[17] Buratti Y, Le Gia Q T, Dick J, Zhu Y and Hameiri Z 2020 *npj Comput. Mater.* **6** 142
[18] Buratti Y, Dick J, Le Gia Q and Hameiri Z 2022 *ACS Appl. Mater. Interfaces* **14** 48647–57
[19] Wang S, Wright B, Zhu Y, Buratti Y and Hameiri Z 2024 *Sol. Energy Mater. Sol. Cells* **277** 113123
[20] Chia J Y, Thamrongsiripak N, Thongphanit S and Nuntawong N 2024 *J. Appl. Phys.* **135** 025701
[21] Olikh O, Lozitsky O and Zavhorodnii O 2022 *Prog. Photovoltaics Res. Appl.* **30** 648–60
[22] Olikh O and Zavhorodnii O 2025 *Sol. Energy* **300** 113754
[23] Haidari G 2025 *Appl. Phys. A* **131** 850
[24] Kim S, Jeong Y, Han D-W and Mo C B 2023 *J. Electron. Mater.* **52** 5861–71
[25] Choudhary K and Sumpter B G 2023 *AIP Adv.* **13** 095109
[26] Kumagai Y 2025 *Materia Japan* **64** 518–23
[27] Ma Y, Yu H, Zhong Y, Chen S, Gong X and Xiang H 2025 *Appl. Phys. Lett.* **126** 044103
[28] Reimer C, Saidi P, Casert C, Beeler C, Tetsassi Feugmo C G, Whitelam S, Mansouri E, Martinez A, Beland L and Tamblyn I 2025 *J. Appl. Phys.* **138** 074306
[29] Tyagi V, Pols M, Brocks G and Tao S 2025 *J. Phys. Chem. Lett.* **16** 5153–9
[30] Mannodi-Kanakkithodi A, Xiang X, Jacoby L, Biegaj R, Dunham S T, Gamelin D R and Chan M K Y 2022 *Patterns* **3** 100450
[31] Mosquera-Lois I, Kavanagh S R, Ganose A M and Walsh A 2024 *npj Comput. Mater.* **10** 121
[32] Olikh O and Zavhorodnii O 2025 *Mater. Sci. Eng. B* **317** 118192
[33] Wang S, Sankaran S and Perdikaris P 2024 *Comput. Methods Appl. Mech. Eng.* **421** 116813

[34] Li W-K and Zhang Y-T 2025 *J. Appl. Phys.* **137** 203304
[35] Kaya M and Hajimirza S 2019 *Sci. Rep.* **9** 5034
[36] Kim Q *et al* 2023 *Solid-State Electron.* **201** 108568
[37] Li X-Y, Sun Q, Xie Y-M and Fung M-K 2024 *Adv. Energy Sust. Res.* **5** 2300263
[38] Haghighat A, Ghadimi A and Eskandarian A 2025 *Plasmonics* **20** 1539–49
[39] Kimerling L and Benton J 1983 *Physica B+C* **116** 297–300
[40] Möller C, Bartel T, Gibaja F and Lauer K 2014 *J. Appl. Phys.* **116** 024503
[41] Olikh O, Kostylyov V, Vlasiuk V, Korkishko R, Olikh Y and Chupryna R 2021 *J. Appl. Phys.* **130** 235703
[42] Li Z, Zhang S, Qu C, Zhang Z and Sun F 2024 *PLoS One* **19** 1–16
[43] Otamendi U, Martinez I, Quartulli M, Olaizola I G, Viles E and Cambarau W 2021 *Sol. Energy* **220** 914–26
[44] Chen X, Karin T and Jain A 2022 *Sol. Energy* **242** 20–29
[45] Munawer Al-Otum H 2024 *Sol. Energy* **278** 112803
[46] Abdelsattar M, Abdelmoety A, Ismeil M A and Emad-Eldeen A 2025 *IEEE Access* **13** 4136–57
[47] Tella H, Hussein A, Rehman S, Liu B, Balghonaim A and Mohandes M 2025 *Case Stud. Thermal Eng.* **66** 105749
[48] Li W-C and Tsai D-M 2012 *Pattern Recogn.* **45** 742–56
[49] dela Rosa M E C, Mateo-Romero H F, Alonso-Gómez V, Ngungu V N, Nava R, Morales Aragonés J I M, Plaza A R, Gonzalez-Rebollo M A, Isaza J R F and Cardeñoso-Payo V 2024 *Renew. Energ.* **2** 27533735241304090
[50] Khanna M, Srinath N K and Mendiratta J K 2020 *Application of Neural Networks and Lifting Wavelet Transform for Long Term Solar Radiation Power Prediction* (Springer) pp 95–105
[51] Held M, Bulling J, Lugovtsova Y and Prager J 2024 *Ultrasonics* **143** 107403
[52] Krishnan S R and Seelamantula C S 2013 *IEEE Trans. Signal Process.* **61** 380–91
[53] Torrence C and Compo G P 1998 *Bull. Am. Meteorol. Soc.* **79** 61–78
[54] Ahmad A, Jin Y, Zhu C, Javed I, Maqsood A and Akram M W 2020 *IET Renew. Power Gener.* **14** 2693–702
[55] Burgelman M, Nollet P and Degrave S 2000 *Thin Solid Films* **361–362** 527–32
[56] Masum Mia M, Faruk Hossain M, Rahman M, Badi N, Irfan A and Ferdous Rahman M 2025 *Mater. Sci. Eng. B* **311** 117817
[57] Joshi T K, Sharma G, Sharma Y R and Verma A S 2024 *Phys. B Condens. Matter* **682** 415793
[58] Ravidas B K, Das A, Agnihotri S K, Pandey R, Madan J, Hossain M K, Roy M K and Samajdar D 2024 *Sol. Energ. Mat. Sol.* **267** 112688
[59] Liu H, Xiang L, Liu Q, Gao P, Zhang Y, Li S and Gao F 2024 *IEEE J. Photovolt.* **14** 59–64
[60] You L, Zhang X, Ma Q, Zhu W and Wu J 2023 *Phys. Status Solidi A* **220** 2300071
[61] Abdulmalik M and Danladi E 2023 *Semicond. Phys. Quantum Electron. Optoelectron.* **26** 321–31
[62] Murphy J D, Bothe K, Olmo M, Voronkov V V and Falster R J 2011 *J. Appl. Phys.* **110** 053713
[63] Wijaranakula W 1993 *J. Etectrochem. Soc.* **140** 275–81
[64] Khelifati N, Laine H S, Vähänissi V, Savin H, Bouamama F Z and Bouhafs D 2019 *Phys Status Solidi a* **216** 1900253
[65] Tan J, Macdonald D, Rougieux F and Cuevas A 2011 *Semicond Sci. Technol.* **26** 055019
[66] Olikh O 2019 *Superlattices Microstruct.* **136** 106309
[67] Fell A *et al* 2015 *IEEE J. Photovolt.* **5** 1250–63
[68] Pässler R 2002 *Phys. Rev. B* **66** 085201
[69] Yan D and Cuevas A 2014 *J. Appl. Phys.* **116** 194505
[70] Green M A 2022 *Prog. Photovoltaics Res. Appl.* **30** 164–79
[71] Couderc R, Amara M and Lemiti M 2014 *J. Appl. Phys.* **115** 093705
[72] Niewelt T *et al* 2022 *Sol. Energy Mater. Sol. Cells* **235** 111467
[73] Black L E and Macdonald D H 2022 *Sol. Energy Mater. Sol. Cells* **234** 111428
[74] Altermatt P P, Schmidt J, Heiser G and Aberle A G 1997 *J. Appl. Phys.* **82** 4938–44
[75] Green M A 1990 *J. Appl. Phys.* **67** 2944–54
[76] Klaassen D 1992 *Solid-State Electron.* **35** 953–9
[77] O'Mara W, Herring R and Hant L 1990 *Handbook of Semiconductor Silicon Technology* (Noyes Publications)
[78] Rougieux F E, Sun C and Macdonald D 2018 *Sol. Energy Mater. Sol. Cells* **187** 263–72
[79] Istratov A A, Hieslmair H and Weber E 1999 *Appl. Phys. A: Mater. Sci. Process.* **69** 13–44
[80] Paudyal B B, McIntosh K R and Macdonald D H 2009 Temperature dependent electron and hole capture cross sections of iron-contaminated boron-doped silicon *2009 34th IEEE Photovoltaic Specialists Conf. (PVSC)* pp 001588–93
[81] Olikh O, Datsenko O and Kondratenko S 2024 *Phys. Status Solidi a* **221** 2400351
[82] Olikh O, Kostylyov V, Vlasiuk V, Korkishko R and Chupryna R 2022 *J. Mater. Sci.: Mater. Electron.* **33** 13133–42
[83] Fadhel S, Delpha C, Diallo D, Bahri I, Migan A, Trabelsi M and Mimouni M 2019 *Sol. Energy* **179** 1–10
[84] Gao W and Wai R-J 2020 *IEEE Access* **8** 159493–510
[85] Srivastava Y and Jain A 2023 *J. Appl. Phys.* **134** 225101
[86] Minagawa H, Tezuka T and Tsuchida H 2024 *Nucl. Instrum. Methods Phys. Res. B* **553** 165383
[87] Lauer K, Möller C, Debbih D, Auge M and Schulze D 2016 Determination of activation energy of the iron acceptor pair association and dissociation reaction *Gettering and Defect Engineering in Semiconductor Technology XVI (Solid State Phenomena)* vol 242 (Trans Tech Publications Ltd) pp 230–5
[88] Zhu X, Yang D, Yu X, He J, Wu Y, Vanhellemont J and Que D 2013 *AIP Adv.* **3** 082124
[89] Zhu X, Yu X, Li X, Wang P and Yang D 2011 *Scr. Mater.* **64** 217–20
[90] Le T T, Zhou Z, Chen A, Yang Z, Rougieux F, Macdonald D and Liu A 2024 *J. Appl. Phys.* **135** 133107
[91] Macdonald D, Roth T, Deenapanray P N K, Bothe K, Pohl P and Schmidt J 2005 *J. Appl. Phys.* **98** 083509
[92] Zhu X, Yu X, Chen P, Liu Y, Vanhellemont J and Yang D 2015 *Int. J. Photoenergy* **2015** 154574
[93] Macdonald D H, Geerligs L J and Azzizi A 2004 *J. Appl. Phys.* **95** 1021–8
[94] Zoth G and Bergholz W 1990 *J. Appl. Phys.* **67** 6764–71
[95] Geerligs L J and Macdonald D 2004 *Appl. Phys. Lett.* **85** 5227–9
[96] Macdonald D, Cuevas A and Geerligs L J 2008 *Appl. Phys. Lett.* **92** 202119
[97] Murphy J D, McGuire R E, Bothe K, Voronkov V V and Falster R J 2014 *J. Appl. Phys.* **116** 053514
[98] Hayamizu Y, Hamaguchi T, Ushio S, Abe T and Shimura F 1991 *J. Appl. Phys.* **69** 3077–81
[99] Schmidt J 2005 *Prog. Photovolt. Res. Appl.* **13** 325–31
[100] Nærland T, Bernardini S, Stoddard N, Good E, Augusto A and Bertoni M 2017 *Energy Proc.* **124** 138–45
[101] (Available at: https://github.com/olegolikh/CV_Fe_SiSC.git)