SiliconPV 2025, 15th International Conference on Crystalline Silicon Photovoltaics
Characterisation, Modelling and Simulation
https://doi.org/10.52825/siliconpv.v3i.2682

Published: 06 Jan. 2026

Correlating Structure Loss and Operational Conditions in Czochralski Silicon Ingot Growth Using Machine Learning

Alfredo Sanchez Garcia[1], Rania Hendawi[2], Hendrik Schön[3], and Marisa Di Sabatino[2]

[1]Sustainable Energy Technology, SINTEF AS, Norway

[2]Department of Material Science and Engineering, Norwegian University of Science and Technology (NTNU), Norway

[3]NorSun, Norway

*Correspondence: Alfredo Sanchez Garcia, alfredo.sanchez@sintef.no

Abstract. This work investigates the relationships between process parameters and the occurrence of structure loss in Czochralski silicon ingots using machine learning. Subsets of features are identified from a dataset of over 14,000 ingots and are used to train random forests to predict structure loss with high accuracy. Multiple rounds of feature importance analysis and refinement are conducted to isolate the process parameters that may have the most significant impact in the occurrence of structure loss. Partial dependence analysis is employed to examine how variations in particular parameters might affect the likelihood of structure loss happening. The results show that the most predictive features of structure loss are primarily recorded late in the process. These features are often influenced by manual interventions or reflect the outcome of structure loss itself. In contrast, early-stage parameters exhibit limited predictive power, suggesting that either early indicators of structure loss are not captured in the available data or that structure loss originates from events occurring later in the growth process. While not predictive in a preventive sense, the model effectively detects deviations from normal operation, thereby demonstrating the value of machine learning for uncovering complex patterns in manufacturing processes data.

Keywords: Machine Learning, Structure Loss, Czochralski Silicon Ingots

1. Introduction

Silicon-based photovoltaics hold 95% of the solar energy market share, with over 96% of these silicon solar cells being monocrystalline due to their higher efficiency and lower defects level [1]. Monocrystalline wafers are sliced from single-crystal silicon ingots produced via the Czochralski (Cz) process [2], a method for silicon growth that has experienced considerable progress in recent years.

The increasing demand for silicon has created a requirement for larger ingot diameters and crucible sizes, presenting challenges in the Czochralski silicon growth method. A considerable number of Czochralski silicon ingots is remelted primarily due to dislocations formed during growth, a phenomenon often referred to as structure loss (SL) [3, 4]. Understanding how SL relates to operational conditions and process parameters during the crystal pulling is crucial, as even small increments in yield can have significant economic benefits. To achieve this, the present work proposes the use of machine learning (ML) to perform a correlation study

and investigate whether certain process parameters are associated with an increased likelihood of SL. The goal of this study is to understand which parameters, or combinations of parameters, contribute to a higher risk of SL. This will ultimately enable better control of the Cz process and reduce the occurrence of structure loss.

This work is structured as follows: section 2 introduces the problem of structure loss during the Cz process. Section 3 presents the workflow and methodology employed in the correlation study. Section 4 highlights the findings obtained with the methodology. Finally, Section 5 gives the main conclusions.

2. Solar Cell Silicon Ingot Production and Structure Loss

The Cz method for single-crystal growth has evolved significantly over the past 50 years and is now a key process for both the photovoltaics and microelectronic industry. The process starts with a seed of a single crystal with a well-defined crystallographic orientation being dipped into the melt and gradually pulled vertically to the surface. The silicon melt solidifies on the seed and adopts its orientation. Precise control of the temperature and the pulling speed is implemented to ensure the formation of a dislocation-free monocrystalline crystal ingot [2, 5]. The main advantage of monocrystalline silicon cells lies in their high efficiency, which is due to the material's purity and low defect density.

One of the crucial challenges of the Cz method is the loss of the dislocation-free structure during growth. This phenomenon is often referred to as SL, and it affects a considerable percentage of the grown ingots at different growth stages. Remelting the affected ingots is the only available solution in the industry, which ultimately decreases the production yield [5, 6]. The root causes for SL occurring during the process are diverse, as illustrated in Figure 1, where three different types of SL are shown. Each example corresponds to a different underlying mechanism. For instance, the structure loss in Figure 1b was triggered by a particle—possibly silicon carbide (SiC)—impacting the growing monocrystalline lattice, leading to a visible transition to multicrystalline silicon.

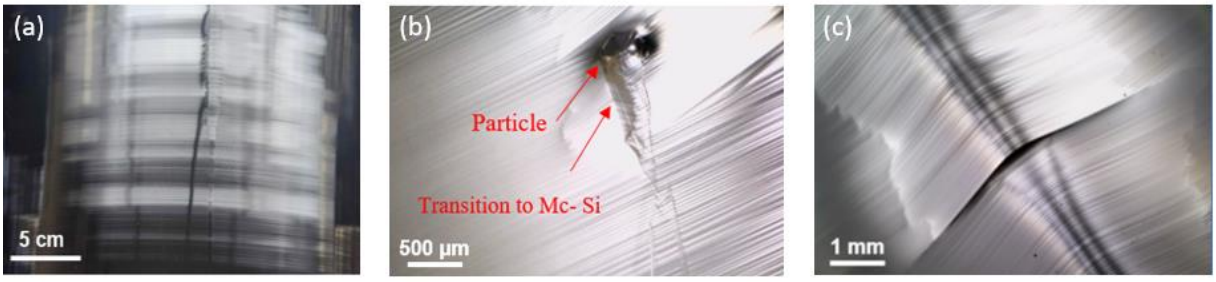

Figure 1. *Optical microscope images of three structure loss categories: (a) diameter fluctuations before the structure loss. (b) Particle hit at the growth front results in nucleation of multicrystalline Si. (c) Cut (notch) in the growth ridge [5].*

3. Machine Learning Methodology for Czochralski Process Analysis

Although the mechanisms by which structure loss appears have been studied, its underlying causes remain to be understood. The Cz process involves a large number of tightly coupled parameters, many of which change dynamically during growth and may interact in nonlinear ways. This complexity makes it difficult to isolate cause-effect relationships using traditional analysis methods. In this context, machine learning (ML) is a promising approach, as it excels at identifying hidden patterns and interactions in high-dimensional datasets without requiring explicit physical models [7]. In particular, ML methods may be used to uncover potential correlations between process parameters and the occurrence of SL, offering a data-driven perspective on this problem.

Figure 2. *Diagram of the machine learning pipeline employed for the correlation study.*

In this work, a subset of ML algorithms, called *classifiers*, was trained on data corresponding to 14000 n-type industrial silicon ingots grown with melt recharging. The data contains process parameters for Cz ingots grown in 20-inch and 22-inch crucibles, covering information of different process parameters, ranging from pulling speeds to the concentration of interstitial oxygen at different ingot positions. A total of 64 process- and material-related variables were utilized as input features for the machine learning model, including both operational settings (e.g., heater power, pulling speed) and measured material properties (e.g., resistivity, oxygen concentration) recorded at two different ingot positions. The hypothesis is that if one parameter, or combination of parameters, is related to the occurrence of SL, then a ML classifier may be able to learn this relationship and distinguish between ingots with and without structure loss. To test this hypothesis, a machine learning pipeline was implemented, as illustrated in Figure 2, and described as follows

1. **Data Grouping:** To preserve confidentiality, all parameter values were first normalized. Datasets were constructed with 64 process parameters plus one column indicating whether SL had occurred or not. SL events were identified through post-growth visual inspection of the silicon ingots by experienced quality control personnel. Then, the datasets were organized into subsets based on logical groupings. For example, process parameters, like pull and crucible speeds, were put together, and the measurement of defect concentrations, such as interstitial oxygen or substitutional carbon, along the ingot, were put together in a different subset. Each subset represents a distinct category of features potentially linked to SL. Randomly sampled subsets were also employed to account for potential non-obvious interactions between Cz process parameters.
2. **Classification with Random Forests:** For each subset, a Random Forest (RF) classifier was trained to predict the occurrence of SL. RF is an ensemble method that builds multiple decision trees on bootstrapped samples of the data, each using a random subset of features [7]. Predictions are made via majority vote. This approach is robust to overfitting, can handle both numerical and categorical data, and provides insight into feature relevance. RF was chosen for its robustness and interpretability, the latter being particularly important for understanding which features drive the model's predictions and may therefore correlate with SL. Under-sampling techniques were used to account for class imbalances.
3. **Feature Importance Analysis:** For subsets showing promising predictive performance, feature importance was evaluated using permutation importance. This technique measures the decrease in predictive accuracy when a feature's values are randomly shuffled [8], thereby estimating its relative contribution to the model's performance. Feature importance analysis allowed for the identification of the most influential features within each subset and to understand which parameters were most strongly associated with SL. It should be noted that permutation importance is context-dependent, as the importance of a feature can change depending on which other features are present.
4. **Partial Dependence Analysis:** The two most influential features of each run were selected for a Partial Dependence analysis to examine how variations in these features might affect the probability of SL, both individually and in combination. A partial dependence plot shows the effect of one or two features on the predicted probability of

SL, and helps visualize threshold behavior, monotonic trends, or interaction effects, although it assumes feature independence.

4. Results

The machine learning classification methodology was run multiple times with different subsets of parameters. The performance of the random forest models is reported using precision, recall, and F1-score metrics, which capture the model's ability to correctly identify structure loss events. Precision refers to the proportion of predicted positives that are correct, recall to the proportion of actual positives correctly identified, and the F1-score is the harmonic mean of precision and recall. Feature importance and interaction effects are also summarized. The parameter names are labeled with P1 and P2, indicating whether they were recorded at the seed-end or at the tail, respectively. Table 1 summarizes the list of parameters used in this section for the correlation study.

Table 1. *Variable names and descriptions of the parameters used for the displayed results.*

Variable Name	Description
P1	Seed end
P2	Tail
CrucibleProducer	Manufacturer of crucible used in the Cz pulling
Resistivity_P2	Resistivity at position 2
BottomHeaterPower_P2	Power of secondary heater at position 2
GasFlow_P2	Gas flow at position 2
MainPressure_P2	Main pressure at position 2
MainPower_P2	Power of main heater at position 2
TestWafer_P1	Exact cut position of characterization wafer (P1)
CrucibleLiftSpeed_P1	Crucible lift speed at position 1
Resisitivity_P1	Resistivity at position 1
CarbonConcentration_P2	Carbon concentration at position 2
ABCD_encoded	Ingot number per run
A/B_encoded	First / last ingot per run
CrucibleLiftSpeed_P2	Crucible lift speed at position 2
SeedSpeed_P2	Seed pulling speed at position 2
MainPressure_P1	Main pressure at position 1

Table 2 shows the results of the RF classifier applied to a subset of parameters containing data on the ingot run (**`ABCD_encoded`**; with values A, B, C and D, representing four runs) and the crucible producer (**`CrucibleProducer`**; with values A and B, representing two different suppliers). The classifier performs better at identifying non-SL events (class 0) but struggles to correctly identify SL (class 1), as indicated by the low precision and F1-score values for class 1. This is understandable given that there is a class imbalance between the two categories in the dataset.

Table 2. *Metrics for each class for the data subset containing the ingot run and crucible types. Here, 0 represents non-SL and 1 represents SL.*

SL	Precision	Recall	F1-Score
0	0.87	0.66	0.75
1	0.16	0.38	0.22

The results presented in Table 2 suggest that there is no correlation between the ingot run and the crucible employed and SL. This may be applicable to the current dataset, given that only two types of crucibles with similar qualities were employed. However, it is counterintuitive, as one would expect that a crucible containing a large number of impurities would contaminate the silicon melt resulting in SL due to particle hit [3]. From this argument, one can conclude that an approach solely based on machine learning may be insufficient to predict structure loss.

Table 3 and Figure 3 show the classification results and feature importance analysis for a model trained on a subset containing two parameters: the resistivity measured at the tail (P2) and the power supplied to the bottom heater at the end of the body growth. The confusion matrix confirms that the model achieves high accuracy in identifying non-SL cases, with a precision of 0.92 and recall of 0.99. Performance on the SL identification is lower, with a precision of 0.86, recall of 0.51, and an F1-score of 0.64. The permutation importance plot indicates that **`Resistivity_P2`** contributes most significantly to the model's predictions, while **`BottomHeaterPower_P2`** plays a smaller role.

Table 3. *Metrics for each class for the data subset containing the resistivity measured at the tail and power from bottom heater measured at the end of the pulling process.*

SL	Precision	Recall	F1-Score
0	0.92	0.99	0.95
1	0.86	0.51	0.64

These results point to a high correlation between structure loss and parameters recorded during the final stage of growth. This trend is general for most parameters recorded at P2. It is worth noting that not all parameters were measured *in situ* during the pulling process. While some parameters, like heater power settings, gas flows and pull speeds, are recorded in real-time, others – such as resistivity – are measured *ex situ* on the grown ingot at room temperature. This distinction is important to consider when interpreting the results, as *ex situ* measurements reflect the final state of the ingot and may be indicative of SL rather than predictive of its occurrence. A straightforward example: significantly smaller ingot weight or length often directly indicate that the pulling process was interrupted prematurely, likely due to SL occurring.

Other P2 parameters carry more indirect information about structure loss. For instance, the resistivity (Figure 3) measured at P2 reflects the doping concentration. Since the resistivity changes along the ingot length according to Scheil's equation, the value of the resistivity implicitly encodes how long the ingot was pulled before termination. This explains why subsets of data containing information on the resistivity at P2 achieve relative high precision.

The pipeline was also applied to a set containing the crucible lift, seed and average pulling speeds. The classification results are shown in Table 4, while the feature importance analysis for these parameters is shown in Figure 4. Table 4 shows that the model performs well at identifying non-SL events (precision = 0.94) but struggles to correctly identify SL events (precision = 0.16). This is the same behavior observed in Table 2 regarding the identification of SL events.

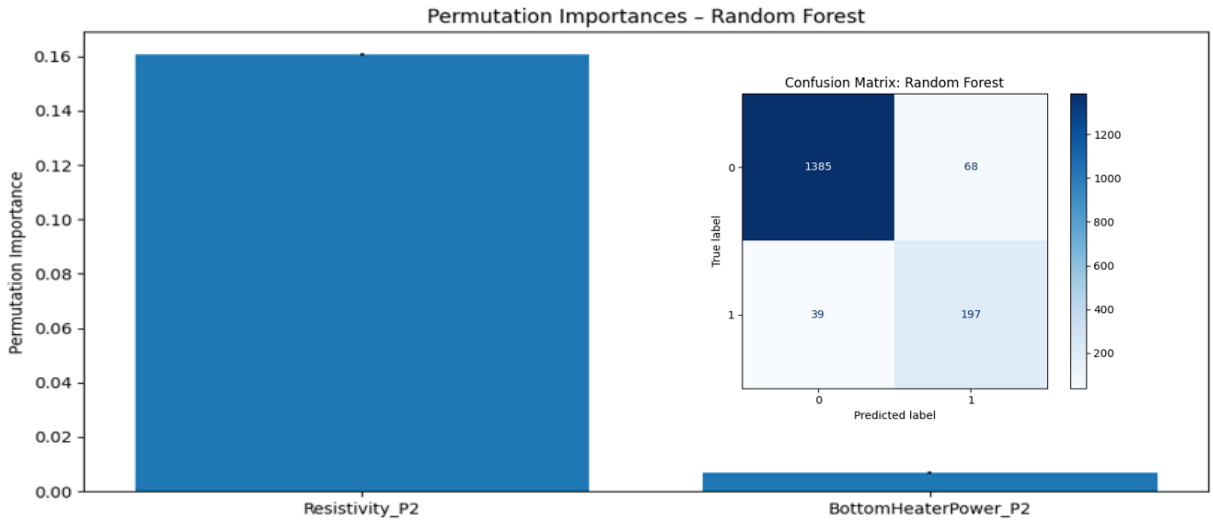

Figure 3. *Model performance and feature importance for predicting structure loss. The confusion matrix shows that the Random Forest classifier achieves high accuracy for the majority class (normal operation). Performance for the minority class (occurrence of SL) is lower. The permutation importance bar plot indicates that* `Resistivity_P2` *contributes most significantly to the model's predictions in this subset, while* `BottomHeaterPower_P2` *plays a comparatively minor role.*

The feature importance analysis, shown in Figure 4, reveals that the crucible lift speed at position 2 (**`CrucibleLiftSpeed_P2`**) contributes most significantly to the model's predictions, followed by the seed speed at position 2 (**`SeedSpeed_P2`**). However, the relatively low overall importance values indicate that these parameters alone do not strongly predict structure loss. Notably, the seed pulling speed at position 1, while included in the model, exhibits a negligible contribution to predictive power. This is again a counterintuitive result, as a faster seed pulling speed at the beginning of the process should correlate with a higher risk of ingot breakage and subsequent structure loss. The lack of a discernible relationship between the seed speed at P1 and the occurrence of SL reinforces the argument that the current methodology may be limited in its ability to capture early-stage indicators of defects, as argued earlier with the results presented in Table 2.

Table 4. *Metrics for each class for the data subset containing the seed, crucible lift and average pulling speeds.*

SL	Precision	Recall	F1-Score
0	0.94	0.24	0.39
1	0.16	0.91	0.28

Figure 5 compares feature importance values derived from subsets of process parameters measured at the start (P1) and end (P2) of the ingot growth process. As noted above, parameters recorded at P2—such as gas flow, chamber pressure, and carbon concentration—demonstrate substantially higher importance in predicting structure loss. However, these parameters likely reflect manual interventions performed after structure loss has occurred. For instance, an operator may adjust heater power upon a visual detection of SL or in an attempt to prevent it. Responses to SL introduce deviations from standard operation, and it appears that the random forest model consistently identifies these anomalies. However, the model does not find a strong correlation between parameters measured at the start of the pull. This suggests that either early indicators of SL are not present from the dataset, or that SL is triggered by events occurring later in the process. Interestingly, Figure 5b shows the carbon concentration at P2 as the most important feature of the employed subset of data, related to SL. This may explain the type of SL mechanism occurring. Due to segregation during solidification, carbon tends to accumulate toward the ingot tail. Elevated carbon concentrations increase the

likelihood of silicon carbide (SiC) particle formation. Such particles can disrupt the monocrystalline lattice—possibly explaining that these instances of structure loss were of the *particle hit* kind. It should be noted, however, that this would normally require carbon concentrations exceeding the solubility limit, and very few of the investigated samples had carbon concentrations above 5 ppma at the tail.

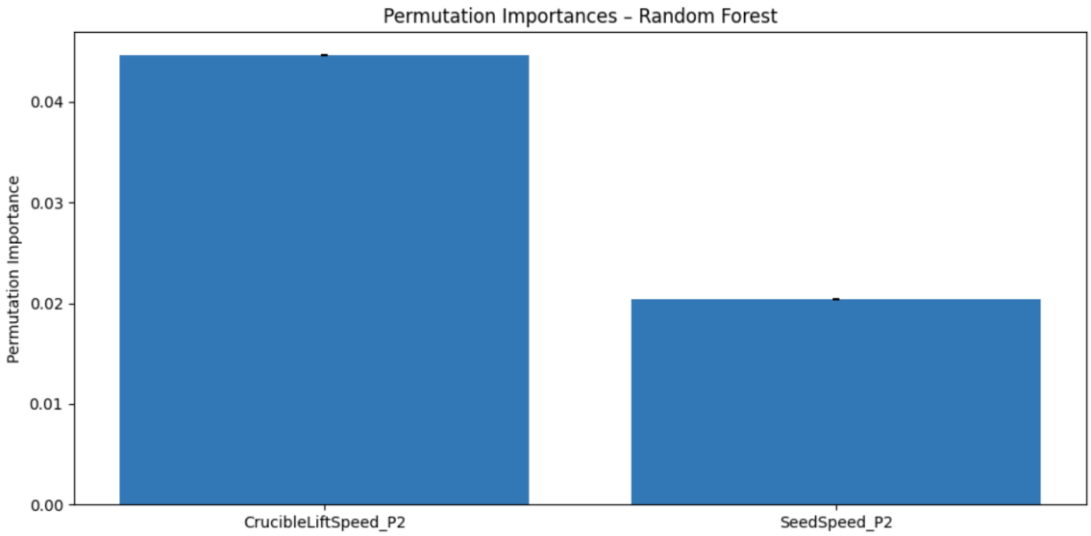

Figure 4. *The permutation importance bar plot for the seed and crucible lift speeds at P2. The results show that* ***`CrucibleLiftSpeed_P2`*** *contributes most significantly to the model's predictions in this subset, with* ***`SeedSpeed_P2`*** *playing a comparatively minor role.*

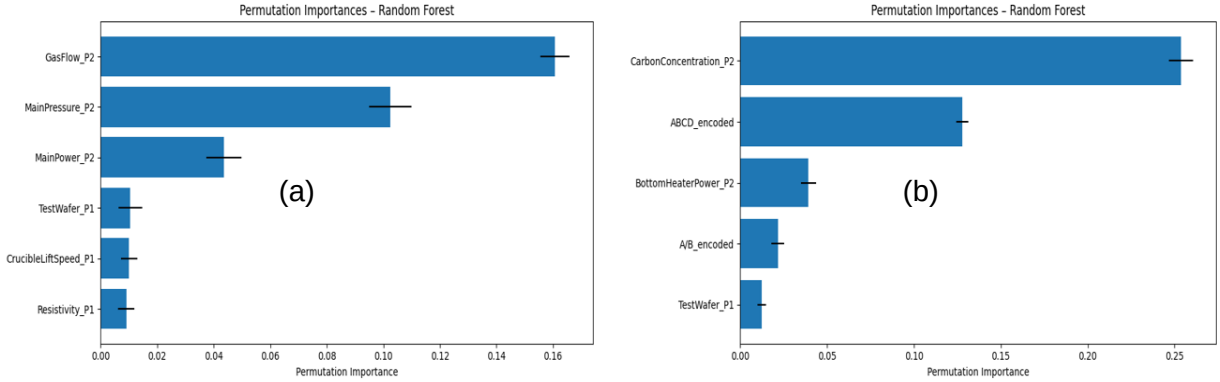

Figure 5. *Feature importance comparison between process parameters recorded at the beginning (P1) and the end (P2) of the ingot production process. Late-stage parameters show significantly higher importance in predicting structure loss.*

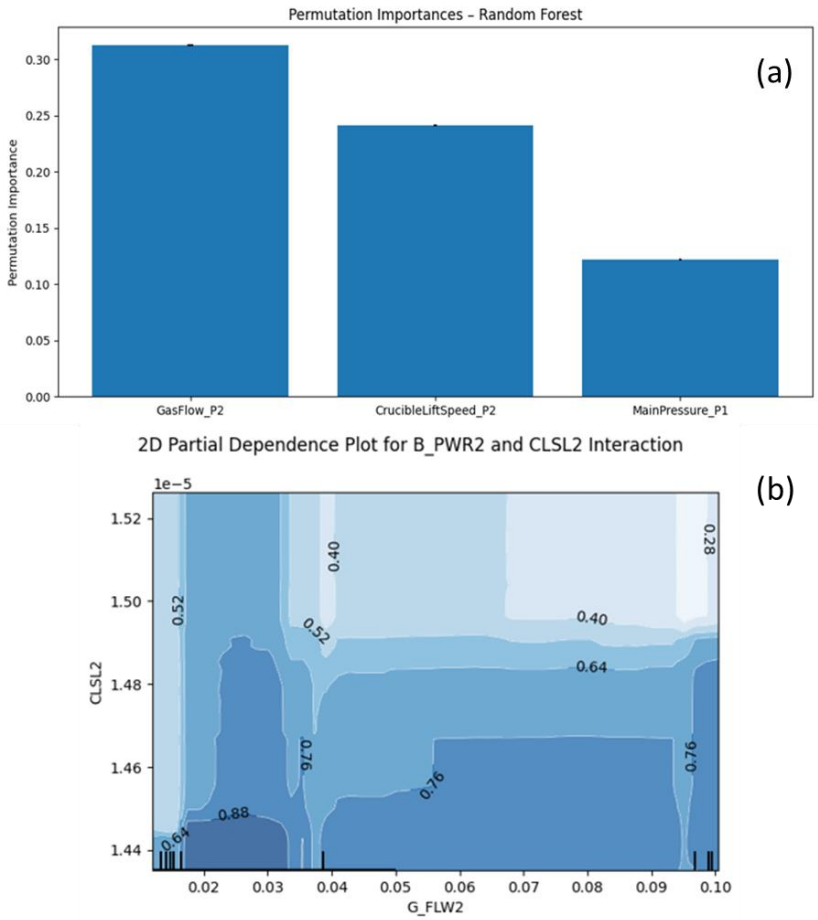

Figure 6. *Feature importance comparison for operational parameters (a, up). Partial dependence plot of the two most influential parameters of this run (b, down): crucible lift speed and gas flow during tailing (P2). The contour lines indicate predicted probability of SL based on those two features. The irregular contour shapes indicate feature interaction.*

Finally, Figure 6a displays feature importance values derived from a subset of process parameters that includes gas flow and crucible lift speed during tailing (P2), as well as main pressure during crowning (P1). The graph highlights the significantly higher importance of the first two parameters compared to main pressure at the beginning of the pulling process. The model interprets that the values of these parameters are outside of what would be considered normal operation. In SL cases, this deviation reflects a different P2 position, i.e., that the pulling process was terminated earlier than expected. Nevertheless, as these two parameters appear to be the most influential of the subset of data, a partial dependence analysis was performed. The goal was to examine how variations in these features might affect the probability of SL. Figure 6b shows the 2D partial dependence plot for gas flow and crucible lift speed, illustrating their combined effect on the predicted probability of structure loss. Recall that the values of these parameters have been normalized to preserve confidentiality. The surface plot does not show a simple relation between the parameters but rather discontinuities, ridges and valleys. This behavior may be linked to abrupt manual intervention, likely by operators, and reinforce the notion that some features recorded at P2 encode reactive changes made after structure loss has already occurred.

5. Conclusions and Outlook

This work has conducted a correlation analysis between the Czochralski pulling process parameters and the occurrence of structure loss using machine learning. The methodology suc-

cessfully identified structure loss events, but most of the predictive capabilities came from parameters recorded after structure loss. Therefore, the model is best suited for *post-process analysis* of structure loss, rather than real-time prediction. In contrast, pulling parameters recorded early in the process show little correlation with structure loss. This suggests that either structure loss is triggered by events occurring later in the pulling, or that early indicators of structure loss are absent from the dataset. One notable limitation of the presented approach is the feature importance analysis being context dependent**.** The importance of a given parameter changes depending on which other parameters are input in the model. See, e.g., the importance of the gas flow parameter during tailing (**`GasFlow_P2`**) in Figures 5 and 6. The importance values are 0.16 and 0.31, respectively. The significant difference in value is due to changes in the feature subset used for training. This reflects that permutation importance is not absolute and, therefore, future work will focus on exploring different techniques within explainable AI, such as *SHapley Additive exPlanations* (SHAP) values, for more robust, interaction-aware interpretation**.**

Data availability statement

The data used in this study were provided by NorSun and are subject to confidentiality agreements; hence, they are not publicly available.

Author contributions

ASG: Conceptualization, Data curation, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. **RH**: Data curation, Writing – review & editing. **HS**: Data curation, Writing – review & editing. **MDS**: Data curation, Writing – review & editing

Competing interests

The authors declare that they have no competing interests

Funding

This work was performed within the Norwegian Research Center for Sustainable Solar Cell Technology (FME SUSOLTECH) and the Norwegian Research Center for Solar Energy (FME SOLAR). The centers are co-sponsored by the Research Council of Norway and their research and industry partners.

References

[1] Fraunhofer Institute for Solar Energy Systems ISE. "Photovoltaics Report—Fraunhofer ISE." *Fraunhofer ISE*. https://www.ise.fraunhofer.de/en/publications/studies/photovoltaics-report.html (accessed Apr. 01, 2024).

[2] Czochralski, J. "Ein neues Verfahren zur Messung der Kristallisationsgeschwindigkeit der Metalle." *Zeitschrift für Physikalische Chemie* 92U, no. 1 (November 1, 1918): 219–21. https://doi.org/10.1515/zpch-1918-9212.

[3] Garcia, A. S., Hendawi, R., & Di Sabatino, M. (2024). Machine Learning Methods for Structure Loss Classification in Czochralski Silicon Ingots. Crystal Growth & Design. https://doi.org/10.1021/acs.cgd.4c00760

[4] Di Sabatino, M., Hendawi, R., & Garcia, A. S. (2024). Silicon Solar Cells: Trends, Manufacturing Challenges, and AI Perspectives. Crystals, 14(2), Article 2. https://doi.org/10.3390/cryst14020167

[5] Hendawi, R., and Di Sabatino M. (2024)“Analyzing Structure Loss in Czochralski Silicon Growth: Root Causes Investigation through Surface Examination.” Journal of Crystal Growth 629 https://doi.org/10.1016/j.jcrysgro.2023.127564.
[6] Hendawi, R., Schön H., and Di Sabatino M. (2025) “Data Analysis of Industrial Czochralski Process: Investigation of Ingots with Structure Loss.” Solar Energy Materials and Solar Cells 283 https://doi.org/10.1016/j.solmat.2025.113438.
[7] MachineLearningMastery.com. “Statistical Methods for Machine Learning.” Accessed April 1, 2025. https://machinelearningmastery.com/statistics_for_machine_learning/.
[8] Breiman, L. “Random Forests.” *Machine Learning* 45, no. 1 (October 1, 2001): 5–32. https://doi.org/10.1023/A:1010933404324.