Lasers in Manufacturing Conference 2023

An intelligent quality inspection system to detect laser welding defects

Patricia M. Dold [a,b,*] , Meiko Boley[a], Fabian Bleier[a], Ralf Mikut[b]

[a]*Bosch Research, Robert Bosch GmbH, Robert-Bosch-Campus 1, 71272 Renningen, Germany*
[b]*Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany*

Abstract

This paper deals with a laser welding process with a high welding speed of 500 mm/s and thin metal plates with a thickness of 75 µm. While the welding process is observed by a photodiode and a high-speed camera in-sito, various in production occurring defects such as spatter, gap or defocus were provoked. To detect welding defects, deep learning has achieved great success and therefore is used as a baseline. However, the results of deep neural networks are difficult to interpret, and their inference times are long. Therefore, our approach empirically extracts and selects relevant features and classifies using decision trees. Results show that our approach leads to competitive results to deep neural networks, but with the advantage of more interpretable results and shorter inference times.

Keywords: laser welding; quality monitoring; machine learning; decision trees; feature importance; signal and image processing

1. Introduction

Laser welding is a key production technology, because of its ability for precise and fast welding. Unfortunately, laser welding processes often are challenging, which leads to welding defects. To quickly detect defects in production, quality monitoring is desired in industrial processes. For quality monitoring different sensors like photodiodes (PD) (Paleocrassas and Tu, 2010), spectrometers (Garcia-Allende et al, 2009), X-ray sensors (Shevchik et al., 2020), optical coherence tomography (Baader et al., 2021) or high-speed cameras (HSC) (Jäger and Hamprecht, 2008) are used. The acquired signals can then be analysed by data-driven methods like support vector machines (You et al., 2014), decision trees (DT) (Hongwei et al., 2011; Moinuddin et al., 2021), random forest algorithms (Wu, 2014) or neural networks (NN) (Zhang et al., 2020). On the one hand, NN have achieved great success in classification tasks; on the other hand, they still have a black box character. In contrast, classical machine learning algorithms like DT provide the advantage of interpretable results. For example, DT can give an importance to features when making a prediction. The important features

could then be interpreted by domain experts or could be used to better understand the data and the DT prediction (Knaak et al., 2018; You et al., 2018).

In the present paper, we classify welding defects based on PD and HSC data. Our main contributions are:

- multi-class classification of laser welding defects and comparison with binary classification in a use case similar to Dold et al., 2022 and Dold et al., 2023,
- usage of classical machine learning including feature engineering and DT and comparison with NN and
- visualization and interpretation of the DT models for better data and model understanding.

2. Data set

The data were acquired in the laboratory. The experimental setup is explained in detail in Dold et al., 2022 and Dold et al., 2023. In short, during the welding process of two thin metal plates with a thickness of 75 µm, photodiode (PD) signals with a sampling rate of 250 kHz and synchronous high-speed camera (HSC) images with a sampling rate of 20 kHz were captured. In total, 59 metal plate pairs (later referred to as experiments) were welded: 9 under reference conditions and 50 with inserted anomalies. While Dold et al., 2023 distinguished between reference and anomaly, the present paper additionally distinguishes between different anomalies. Figure 1 shows schematically the different categories and the captured sensor data. Thereby, the green box includes the reference category and the red box the anomaly categories. The welding direction in the shown HSC images is from bottom to top. The PD provides voltages V over time t. Because of the different sampling rates of the sensors, while one HSC image is acquired, there are several PD voltages. Several in production occurring cases were readjusted:

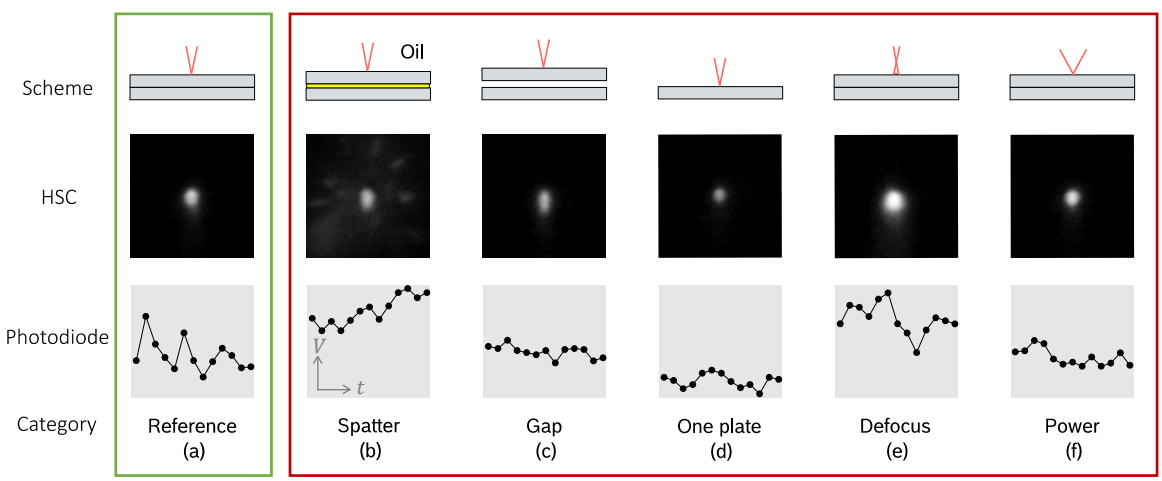

Fig. 1. Six different categories have been considered: green box: reference welding (a); red box: different anomalies have been introduced in the laser welding process to provoke defects. In (b) oil was inserted between the plates, which led to spatters. In (c) a gap between the plates was inserted. (d) was welded with one plate only, in (e) the laser was defocused and in (f) the laser power was changed.

Table 1. Number of chunks of the data set. Given are the average numbers over 5 folds of a cross-validation.

Category	n_{train}	n_{test}	n
Reference	148425	37106	185531
Spatter	25051	6263	31314
Gap	30895	7724	38619
One plate	27044	6761	33805
Defocus	37862	9465	47327
Power	26516	6629	33145
Total	295793	73948	369741

(a) For the reference welding, the metal plates were perfectly positioned on top of each other. In the HSC image the keyhole is visible.
(b) A thin film of oil was applied between the two plates to simulate contaminations. Because of the high temperatures of the laser, the oil expanded and evaporated. Therefore, the material of the plates ejected, so the process spattered.
(c) A gap was inserted between the two plates. In the HSC image a more widely opened capillary is visible. The standard deviation of the PD signal compared the reference category fell.
(d) Only one plate was welded. Thereby, the workpiece was welded through, which led to a lower light reflection. Therefore, the intensity values in the HSC image and the voltages in the PD signal decreased.
(e) The laser beam hit the workpiece defocused. Compared to the reference category, the HSC images show a bigger capillary radius, and the PD signals show higher voltages.
(f) The laser power was changed.

The number of chunks of the categories are given in Table 1. A chunk consists of one HSC image and 13 corresponding PD samples. The average number of training chunks is given by n_{train}; the average number of test chunks by n_{test} and both together as n.

3. Classification approaches

To distinguish between the introduced categories, different classification approaches have been applied. Firstly, features of the PD and HSC data were extracted. After that, the extracted features were used for classification with decision trees (DT). As a comparison, a deep learning approach consisting of a neural network (NN) was used. All models were implemented in python. Moreover, for all models a 5-fold cross-validation was applied.

3.1. Feature engineering

Features of the PD signals were extracted in two ways: manually (7 features) and automated (794 features) with the python toolbox tsfresh (Christ et al., 2018). A detailed description of the feature extraction can be found in Dold et al., 2023. For the HSC images, statistical (8 features) and geometrical (165 features) features were selected. The geometrical features were calculated depending on a threshold $h \in [0, 255]$. Thereby, two different types of thresholds h, namely absolute thresholds h_a that are the same for each image, and thresholds h_q based on quantiles where the value can differ for each image. Based on the threshold, a binary mask image was calculated for every HSC image. The mask has the value 0 where the pixel values in the HSC

images are smaller than h, and 1 where the pixel values are bigger or equal. Based on the mask, the following features were extracted: area, number of regions, area of the biggest region, ratio of area of the biggest region and area, convex hull, ratio of area and convex hull, circumference, ratio of circumference and area, area of a fitted ellipse, length of the ellipse and width of the ellipse. Figure 2 visualizes the features at $h_a = 40$ for the categories reference, spatter, and gap. In (a) the original images, already known from Figure 1, are shown. (b) shows the binary mask, from which the area and the number of regions were derived. (c) shows the biggest region of the mask and (d) the circumference of the mask. (e) shows the convex hull, which mainly differs from the circumference when there are spatters. In (f) the ellipse fit of the biggest region is shown, from which the length l and the width w of the ellipse were derived. Differences in the features depending on the category are visible, e.g. the number of regions and the convex hull increase for a spattering process.

3.2. Decision trees

After feature extraction, decision trees (DT) were used for classification. The DT were implemented with the python library scikit-learn (Pedregosa et al., 2018). Hyperparameters of the DT, namely the maximum depth of the tree and the minimum number of samples required to split an internal node, were found with a grid search. The importance of each feature in the DT was calculated to reduce the number of input features. To build the final DT the features that were at least under the ten most relevant features in one of the trees of the 5-fold cross-validation were chosen.

3.3. Convolutional neural networks

Besides feature engineering with following DT classification, the HSC and PD data were classified with convolution neural networks. For the PD data CNN1 and for the HSC images RN50 based on ResNet (Ke et al.,

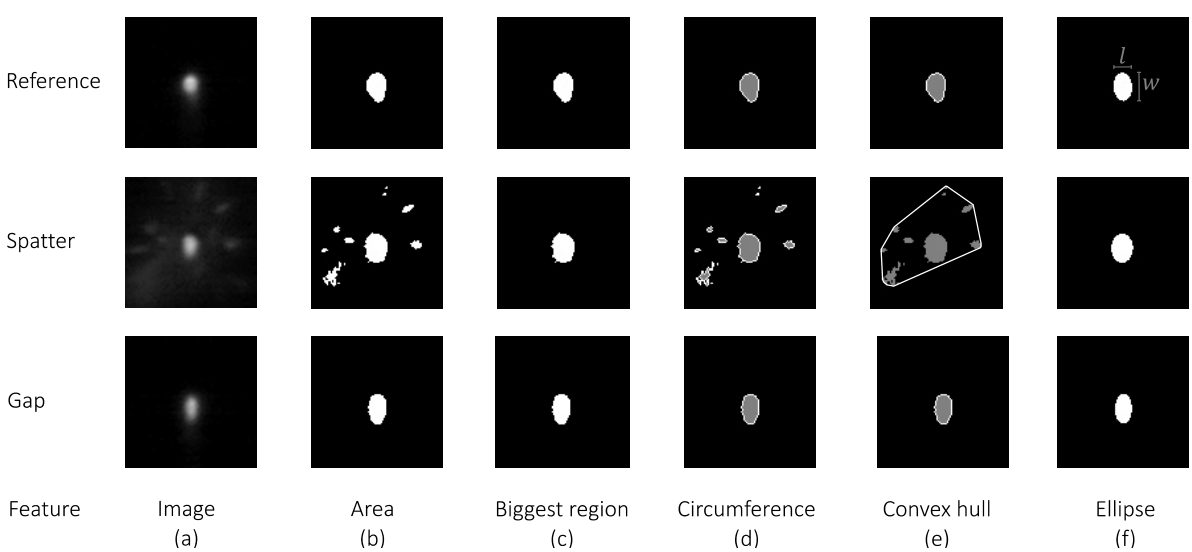

Fig. 2. Geometrical feature extraction of the HSC images at $h_a = 40$ for the three categories reference (first row), spatter (second row) and gap (third row). (a) shows the original image. In (b) the binary mask is visualized. (c) shows the biggest region of the binary mask, (d) the circumference of the mask and (e) the convex hull of the mask. In (f) the ellipse fit of the biggest region is shown, from which the length l and the width w were derived.

2016) was used. The concrete architectures and training parameters of CNN1 and RN50 are described in Dold et al., 2023. However, instead of the sigmoid activation function, a softmax layer is used at the output due to multi-class classification.

4. Results and discussion

The classification results based on feature extraction with DT are compared with results from NN. To understand a DT model's prediction, the importances of the features, of which a DT is build, are calculated. This is rather difficult for deep learning approaches consisting of NN. The NN classification consists of several thousands of computational operators and, therefore, it is difficult to determine which parts of an input sample lead to a certain decision.

4.1. Multi-class and binary classification

Table 2 shows the classification results, namely accuracy, precision, recall and F1-score of different models based on the PD and the HSC data. Thereby, the mean over the 5 folds of the cross-validation is given. To calculate the metrices in the multi-class case, micro averaging was used. Table 2 compares 1) multi-class with binary classification, 2) experiment split (∗) with random split (+), and 3) DT with NN. Thereby, binary classification refers to only having the two categories reference and anomaly (see Fig. 1 green and red boxes) and multi-class to all six categories. Besides a training and test split based on the welding experiments (∗), as described in Dold et al., 2023, a random split (+) of the samples was implemented. The split according to experiments (∗) is closer to the production scenario as algorithms are trained on data of some workpieces and then applied to data of other ones. However, a random split (+) leads to a more similar distribution of the training and test data.

For the PD DT models (I and II), the first line of the evaluation metrices refers to the features extracted from tsfresh and the second line to the seven manually extracted features. The accuracy differs by 0.48% for binary classification and 1.82% for multi-class classification. As the results are competitive, the manually extracted features already capture most information to distinguish between the categories. Comparing the PD results (I and II), the multi-class classification performs 9.00% (tsfresh features) and 10.34% (manually features) worse

Table 2. Classification results averaged over the 5 folds of the cross-validation. Given are accuracy, precision, recall and F1-score for the different models, which were created based on the PD and HSC data. (∗) indicates a split of training and test set based on experiments and (+) indicates a random split independent of experiments. For the PD models DT binary (∗) and DT multi (∗), the first line of the evaluation metrices refers to the features extracted from tsfresh and the second line to the seven manually extracted features.

		PD				HSC			
Number	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
I	DT binary (∗)	88.18 87.70	88.31 86.94	88.03 88.79	88.14 87.80	93.81	91.43	96.88	94.05
II	DT multi (∗)	79.18 77.36	79.18 77.36	79.18 77.36	79.18 77.36	89.50	89.50	89.50	89.50
III	DT binary (+)	82.97	76.11	89.51	82.27	94.13	91.76	97.03	94.32
IV	DT multi (+)	80.35	80.35	80.35	80.35	90.58	90.58	90.58	90.58
V	NN binary (∗)	89.84	86.86	94.02	90.21	96.50	95.13	98.18	96.61
VI	NN multi (∗)	79.42	79.42	79.42	79.42	92.90	92.90	92.90	92.90

True

Predicted	(b)	(c)	(d)	(e)	(f)	∑
(b)	97	3287	263	245	348	312
(c)	602	532	4369	1105	18	842
(d)	1	462	1343	5403	0	12
(e)	258	651	8	0	7630	73
(f)	892	454	626	8	197	2604
∑	37106	6263	7724	6761	9465	6629

True

Predicted	(b)	(c)	(d)	(e)	(f)	∑
(b)	177	4265	371	149	324	219
(c)	63	321	6642	104	4	144
(d)	0	218	173	6503	0	15
(e)	322	452	3	0	7968	60
(f)	369	182	134	5	53	4636
∑	37106	6263	7724	6761	9465	6629

Table 3. Confusion matrices of multi-class classification with DT. The left confusion matrix results from a model based on PD data and the right confusion matrix based on HSC data (see II of Tab. 2). The categories are (a) reference, (b) spatter, (c) gap, (d) one plate, (e) defocus and (f) power change (see Fig. 1). Given are the averages over the 5 folds of the cross-validation.

than the binary classification. Regarding the different splits, the random split (+) slightly outperforms the experiment split (⊛) for the multi-class case but performs worse for the binary case. The NN based models (V and VI) slightly outperform the DT; but both DT and NN are competitive.

For the HSC models (I and II), binary classification leads to an accuracy of 93.81% and multi-class classification to 89.50%. Compared with the results of the PD models, the HSC models perform 5.63% better for the binary case and 10.32% better for the multi-class case. The random split (+) slightly outperforms the experiment split (⊛). As for the models based on PD data, the NN perform better but the results still are competitive.

Table 3 shows the confusion matrices of multi-class classification with DT. The left matrix results from the model based on PD data and the right confusion matrix based on HSC data (see II of Tab. 2). The categories are (a) reference, (b) spatter, (c) gap, (d) one plate, (e) defocus, and (f) power (see Fig. 1). Given are the averages over the 5 folds of the cross-validation. The left matrix shows that the model misclassifies some gap, defocus and power samples as reference. Furthermore, the model has difficulties to distinguish gap from one plate. In contrast, there is almost no misclassification of reference and one plate, one plate and defocus, one plate and power, or gap and defocus. The right matrix shows that the model based on HSC data also misclassifies some defocus or power samples as reference. However, compared with the left matrix, there are fewer misclassifications.

4.2. Feature importances and interpretation of decision trees

To better understand a DT model's prediction, evaluation of the importances of the features is useful. The feature importances give a score to each feature of which a DT model is built. So, the scores give the importance of each feature when making a prediction. In the following, the feature importance is defined mathematically. Therefore, first the Gini impurity P at a node is defined as

$$P = 1 - \sum_{c \in C} (s_c)^2 \in [0,0.5], \qquad (1)$$

with the categories C and the proportion of samples s_c of category c at the node. The impurity reduction R_m at node m with two child nodes then is

$$R_m = P_m - \left(\frac{N_1}{N_m} P_1 + \frac{N_2}{N_m} P_2\right), \qquad (2)$$

with the Gini impurity P_m at node m and the Gini impurities $P_1$and P_2 at the two child nodes of m. N_m is the number of samples at node m and N_1 and N_2 the sample numbers at the child nodes, respectively, so $N_m = N_1 + N_2$. With the features G, the current feature g_i and the nodes M_{g_i}, which perform a split on feature g_i, the feature importance I_{g_i} of the feature g_i of a decision tree is,

$$I_{g_i} = \frac{\sum_{m \in M_{g_i}} R_m}{\sum_{g \in G} \sum_{m \in M_g} R_m} \in [0,1]. \qquad (3)$$

Because of the normalization in the denominator, the feature importances of all features add up to 1. Figure 3 shows the interpretation of DT for multi-class classification of PD and HSC data. Firstly, in Figure 3 I the feature importances of the most important features are given. The features are given on the x-axis and the importances on the y-axis. The importances of the 5 individual folds are drawn in blue and their mean in red. For the PD data, the most important features include maximum, standard deviation, mean and minimum, which are four out of the seven manually extracted features. This is consistent with the results from Table 2, where the DT based on the seven manually extracted features led to competitive results to the DT based on automated extraction with tsfresh. Moreover, the root mean square has the highest importance. The exact meaning of the other relevant features can be looked up in Christ et al., 2018. It is noticeable that there is a strong scatter in terms of the importances between the folds. In contrast, for the HSC data, there is less scatter between the folds. For the HSC data, three features clearly have the highest importance (area, $h_a = 85$; ellipse length $l, h_a = 10$; ellipse width $w, h_a = 85$).

Next, Figure 3 II shows the root node and one child of the DT of the first fold based on PD or HSC data. Thereby, the vector $p = [a, b, c, d, e, f]$ gives the percentage of samples of each category (a)-(f) (see Fig. 1). So, for example the first entry a gives the percentage of samples of category (a), which is reference, in the considered node in relation to the total available reference samples. Besides the vector p, the second entry in the nodes of the DT is the split criterion. In the DT on the left based on PD data, all entries of the p vector are 100% as all samples are available at the beginning. The split in the root node is based on the root mean square, which has the highest feature importance after I. The child node then has the absolute maximum as criterion, which has the third highest importance after I. On the right side, there is the DT based on HSC data. The split feature in the root node is ellipse width $w, h_a = 85$, which is the third important feature after I. For the child node the split feature is ellipse length $l, h_a = 10$, which is the second important feature after I.

Finally, Figure 3 III shows the feature in the child node over the feature in the root node from II. The feature values of each sample are plotted in the color of their category and different clusters are formed. Additionally, the decision boundaries are marked in gray. In the left plot based on the PD data, the vertical line at root mean square $= 0.043\ V$ separates category defocus from reference, gap, one plate and power. The entries of p in the child node in II confirm that: 100% of the samples of the categories reference, gap, one plate and power are in that node. Moreover, 85% of the defocus samples are eliminated. The horizontal line at absolute maximum $= 0.036\ V$, so the criterion of the child node, splits the remaining samples further: reference and defocus are separated from gap, one plate and power. In the right plot based on the HSC data the criterion of the root node separates reference, defocus and power from gap and one element. This is also indicated by the

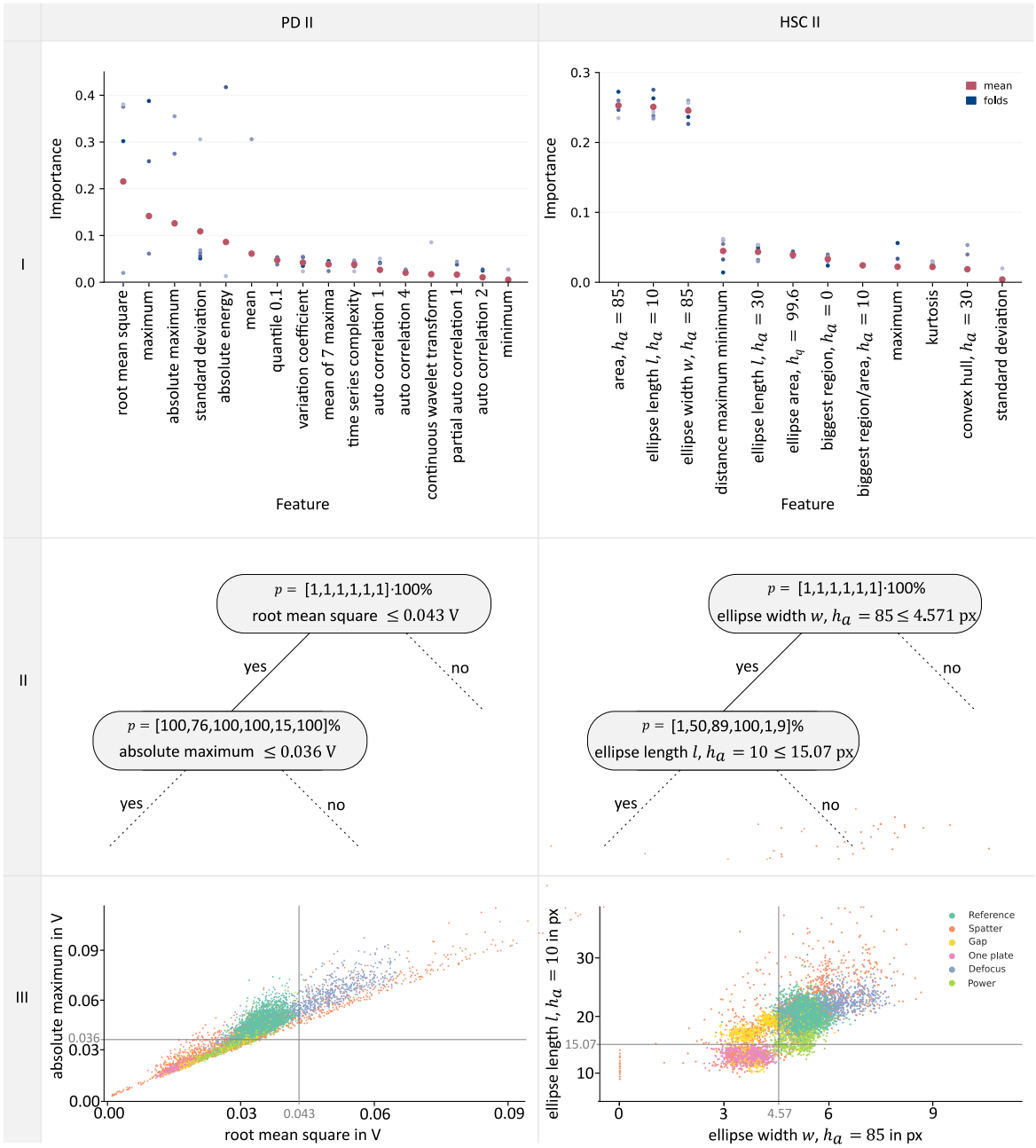

Figure 3. Interpretation of the DT for multi-class classification based on PD and HSC data. I shows the most important features with their feature importances. II shows the root node and one child node of the DT of the first fold. Thereby, the vector $p = [a, b, c, d, e, f]$ gives the percentages of samples of each category (a)-(f). III shows the feature of the child node over the feature of the root node from II. The feature values of each sample are plotted in the color of their category. Additionally, the decision boundaries are marked in gray.

p vector. The criterion of the child node then separates gap from one plate. When comparing the PD and the HSC plots based on two relevant features, the HSC plot contains better separable clusters. For examples, the categories gap and power are hardly distinguishable in the PD scatter plot but in the HSC plot they build separable clusters.

4.3. Combining photodiode and high-speed camera classifiers

So far, the models based on PD and HSC data were considered separately. Both, models based on PD and HSC data, have their advantages and disadvantages: As shown in Table 2, the PD models have lower accuracies than the HSC models. However, as for the PD only 13 samples must be analysed compared with a whole image of the HSC, evaluation times are faster. As welding processing become faster, there is a need for fast but still precise quality inspection, so the advantages of the PD and HSC evaluation should be used together. Therefore, Dold et al., 2022 proposed a two-stage quality monitoring system that first analyses the PD data and only in case of uncertainty of that model's decision, an analysis of the HSC data is considered. For the DT in the present paper, the uncertainty could be given by the fraction of samples belonging to the same category in a leaf. Dold et al., 2022 and Dold et al., 2023 showed that a combined classification system can reduce the computational effort significantly while the accuracy stays stable.

5. Conclusion

In the present paper, we present classification algorithms, namely decision trees (DT) and neural networks (NN) to distinguish welding defects based on PD and HSC data. Thereby, multi-class defect classification was compared with binary classification. The multi-class classification performs about 10% better on the HSC than on the PD data. The NN slightly outperform the DT but both approaches still lead to competitive results. Moreover, DT have the advantage of being easier to interpret. By calculation the feature importances of each feature in a DT, the most important features for prediction were found: For the PD data simple statistical features are important. For the HSC data the area of the keyhole and its ellipse length and width are relevant.

References

Baader, M., Mayr, A., Raffin, T., Selzam, J., Kühl, A., & Franke, J. (2021). Potentials of Optical Coherence Tomography for Process Monitoring in Laser Welding of Hairpin Windings. *2021 11th International Electric Drives Production Conference (EDPC)*,1-10.

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfesh–a python package). *Neurocomputing*, 307, 72-77

Dold, P. M., Bleier, F., Boley, M., & Mikut, R. (2022). Multi-stage Inspection of Laser Welding Defects using Machine Learning. *Proceedings 32. Workshop Computational Intelligence*, 1, 31.

Dold, P. M., Bleier, F., Boley, M., & Mikut, R. (2023). Two-stage Quality Monitoring of a Laser Welding Process Using Machine Learning. (s*ubmitted)*.

Garcia-Allende, P. B., Mirapeix, J., Conde, O. M., Cobo, A., & Lopez-Higuera, J. M. (2009). Spectral processing technique based on feature selection and artificial neural networks for arc-welding quality monitoring. *Ndt & E International*, *42*(1), 56-63.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.

Hongwei, X., Xianmin, Z., Yongcong, K., & Gaofei, O. (2011). Solder joint inspection method for chip component using improved AdaBoost and decision tree. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, *1*(12), 2018-2027.

Jäger, M., & Hamprecht, F. A. (2008). Principal component imagery for the quality monitoring of dynamic laser welding processes. *IEEE Transactions on Industrial Electronics*, *56*(4), 1307-1313.

Knaak, C., Thombansen, U., Abels, P., & Kröger, M. (2018). Machine learning as a comparative tool to determine the relevance of signal features in laser welding. *Procedia CIRP*, *74*, 623-627.

Moinuddin, S. Q., Hameed, S. S., Dewangan, A. K., Kumar, K. R., & Kumari, A. S. (2021). A study on weld defects classification in gas metal arc welding process using machine learning techniques. *Materials Today: Proceedings*, *43*, 623-628.

Paleocrassas, A. G., & Tu, J. F. (2010). Inherent instability investigation for low speed laser welding of aluminum using a single-mode fiber laser. *Journal of Materials Processing Technology*, *210*(10), 1411-1418.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Shevchik, S., Le-Quang, T., Meylan, B., Farahani, F. V., Olbinado, M. P., Rack, A., Masinelli, G., Leinenbach, C., & Wasmer, K. (2020). Supervised deep learning for real-time quality monitoring of laser welding with X-ray radiographic guidance. *Scientific Reports*, *10*(1), 1-12.

Wu, H. (2017). Solder joint defect classification based on ensemble learning. *Soldering & Surface Mount Technology*, *29*(3), 164-170.

You, D., Gao, X., & Katayama, S. (2014). Multisensor fusion system for monitoring high-power disk laser welding using support vector machine. *IEEE Transactions on Industrial Informatics*, *10*(2), 1285-1295.

You, D., Gao, X., & Katayama, S. (2016). Data-driven based analyzing and modeling of MIMO laser welding process by integration of six advanced sensors. *The International Journal of Advanced Manufacturing Technology*, *82*, 1127-1139.

Zhang, Y., You, D., Gao, X., Wang, C., Li, Y., & Gao, P. P. (2020). Real-time monitoring of high-power disk laser welding statuses based on deep learning framework. *Journal of Intelligent Manufacturing*, *31*, 799-814.