Enhancing manual revision in manufacturing with AI-based defect hints

Jože M. Rožanec, Patrik Zajec
Jožef Stefan International Postgraduate School
Jamova cesta 39, 1000 Ljubljana, Slovenia
{joze.rozanec, patrik.zajec}@ijs.si

Jelle Keizer
Philips Consumer Lifestyle BV
Faculty of Electrical Engineering and Computer Science
Oliemolenstraat 5, Drachten, The Netherlands
jelle.keizer@philips.com

Elena Trajkova
University of Ljubljana
Faculty of Electrical Engineering
Tržaška 25, Ljubljana, Slovenija
trajkova.elena.00@gmail.com

Blaž Fortuna
Qlector d.o.o.
Rovšnikova 7, Ljubljana, Slovenija
blaz.fortuna@qlector.com

Bor Brecelj, Beno Šircelj, Dunja Mladenić
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
bor.brecelj@gmail.com, {beno.sircelj, dunja.mladenic}@ijs.si

Abstract. *Quality control allows companies to verify whether the products conform to requirements and specifications. However, while Artificial Intelligence is increasingly used to automate the visual inspection process, a manual revision can be required when the model cannot determine whether a piece is defective or not with enough confidence. Therefore, means must be devised to optimize the manual revision of such products, to increase the speed and quality of labeling. In this paper, we perform experiments to determine whether different defect hinting techniques and data imbalance mitigation techniques can enhance the manual revision process. Furthermore, we contrast the performance of two groups of persons with different skills and education levels and their perceptions when executing the experiments. We performed the experiments on real-world data provided by Philips Consumer Lifestyle BV.*

Keywords. Intelligent Manufacturing Systems; Artificial Intelligence; Quality Assurance and Maintenance; Fault Detection; Human Centred Automation

1 Introduction

Quality control is one of the critical activities performed in the manufacturing industry to detect product defects to ensure quality standards are met, avoid rework, and supply chain disruptions, and avoid potential damage to the brands' reputation (Wuest et al. (2014); Yang et al. (2020)). In addition, information regarding the defective products enables precise and timely root cause detection and therefore considers mitigation actions to improve the manufacturing processes and overall product quality.

The advent of the Industry 4.0 paradigm has fostered the use and integration of information and communication technologies to increase the flexibility and efficiency of the manufacturing process while also leading to greater value over the whole product lifecycle (Frank et al. (2019)). The decreased cost of sensors has accelerated the digitalization of manufacturing (Benbarrad et al. (2021)), enabling the use of artificial intelligence for defect detection in industrial settings (Carvajal Soto et al. (2019); Chouchene et al. (2020)). Automated visual inspection provides greater scalability (does not require training and is not subject to fatigue, absenteeism, or inspection inefficiencies as humans are) and ensures all products are inspected following the same criteria. Given the abovementioned benefits, machine learning automated visual inspection has been applied in many scenarios (Gobert et al. (2018); Iglesias et al. (2018)). While supervised classification is frequently used to discriminate between known defects, unsupervised machine learning approaches enable discovering non-labeled defects or performing defect detection where such labeling is not feasible.

Regardless of the machine learning approach used to perform the automated visual inspection, there are usually certain products for which the machine learning model cannot determine with a high level of certainty whether they are defective or not. In such cases, manual inspection is required to perform a final evaluation of the product's quality. Therefore, we consider

it essential to develop means to enhance the manual revision process to achieve an increased speed and quality of labeling. This research compares balanced and imbalanced data streams with defect hinting strategies to understand which yield better outputs and how the users perceive them.

The main contribution of this research is the assessment of multiple scenarios regarding manual revision involving balanced and imbalanced data streams and different defect hinting strategies. The scenarios are assessed based on data collected during the experiments to determine the time and quality of labeling and the participants' perceptions evaluated through a set of surveys. The experiments were performed considering a real-world use case from the *Philips Consumer Lifestyle BV* corporation.

We evaluate the labeling quality through the precision, recall, and F1 metrics, considering it a binary classification problem (whether a defect exists or not, regardless the defect). Furthermore, we compare the mean and median labeling times and compute the labeling time standard deviation. Finally, we quantify the number of unidentified defects. To understand the participants' perception, we summarize the survey results and compare the outcomes obtained for two different groups: (a) students and researchers of an artificial intelligence laboratory and (b) *Philips Consumer Lifestyle BV* operators.

The rest of this paper is structured as follows: Section 2 presents related work, and the Section 3 describes the *Philips Consumer Lifestyle BV* use case. Section 4 introduces the methodology we followed in executing the experiments later. Section 5 describes the experiments we performed, while Section 6 describes and analyzes the results we obtained. Finally, Section 7 offers our conclusions and provides an outline for future work.

2 Related Work

Automated visual inspection is implemented with image processing techniques. Among them, machine learning models had been widely adopted, and state-of-the-art (SOTA) performance was achieved with deep learning models (Beltran-Gonzalez et al. (2020); Pouyanfar et al. (2018)). While such models achieve great precision, when they cannot classify a product with high confidence, the product must be manually inspected or thrown away. EXplainable Artificial Intelligence (XAI) can provide additional insights on such cases to understand gaps in the model's knowledge and improve it over time. Nevertheless, little research was devoted to the problem of manual revision when such products are manually inspected. Furthermore, there is a research void on how cues provided by XAI techniques or unsupervised machine learning models can assist the operators in manual revision.

The heatmap is a data visualization technique that aims to characterize a phenomenon through a color scale extended over two dimensions. As such, it has been widely adopted by XAI techniques related to image classification (Samek and Müller (2019)), and unsupervised machine learning models (Defard et al. (2021); Zavrtanik et al. (2021)), to convey information regarding what is being considered by the algorithm to make a prediction or where does the model expect the defect to be, frequently overlaying the heatmap with a certain degree of transparency to the original image.

3 Use Case

The research was performed based on a real-world use case concerning the visual inspection of manufactured products produced by *Philips Consumer Lifestyle BV* in Drachten, The Netherlands. The manufacturing plant is considered one of the largest Philips development and production centers in Europe. Printing machine setups for various products and logos exist, and many products produced on these machines are manually handled and inspected to determine their visual quality. If some defect is observed, they are removed from the manufacturing line, leaving only those manufactured products that comply with the existing quality standards. Operators spend several seconds handling, inspecting, and labeling the products. While such a process could be automated to a certain extent, there are certain cases whose quality cannot be automatically determined and therefore require manual inspection. Therefore, we explore means to increase the operators' labeling speed and labeling quality through artificial intelligence. We do so through the creation of synthetic images to achieve class balance and by providing defect hints using heatmaps obtained either from XAI techniques (GradCAM (Selvaraju et al. (2017))) or unsupervised defect detection models (DRAEM (Zavrtanik et al. (2021))), or showing the nearest labeled images. The research was performed with a labeled dataset of 3.518 images, with three possible categories (see Fig. 1): good printing, double printing, and interrupted printing.

Figure 1: The images considered for this research correspond to well-printed logos (good) or defective prints (double or interrupted prints).

4 Methodology

For this research, we developed a web application that enabled the participants to log in with an assigned user-

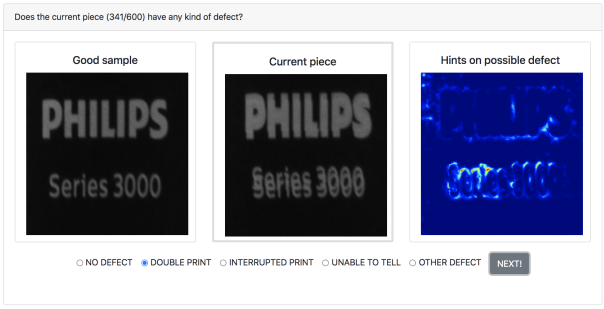

Figure 2: Screenshot of an application we developed to test multiple labeling scenarios. We always display a *good* image on the left side. The image to be inspected is shown in the center. On the right, we eventually provide some defect hinting. In the sample screenshot, we find an image of a *double-print* defect and the hint image obtained from a DRAEM anomaly map.

name and password and sequentially execute a series of experiments. When a given experiment was completed, the next one was unlocked to ensure all participants executed the experiments in a particular order. While the information collected during the experiments was associated with the username of the corresponding participant, no means were provided to associate the usernames back to a particular person. For each experiment, we collected the experiments' start and end time, the time required to label each image, and the labels the participants assigned to each image. For each image, we had prior information regarding the ground-truth label and whether the image was obtained from the original dataset provided by *Philips* or if it was synthetically generated.

We asked the participants to execute one experiment daily to ensure the labeling fatigue would not affect their performance. In addition, we asked them to fill out two kinds of surveys: (a) surveys to understand their perception regarding a particular experiment setup (Experiments 3-6), and (b) an overall survey to obtain basic demographic data, evaluate the overall experience, and compare the perceived usefulness of the setups suggested by each experiment, once all of them were completed. In the first survey, we asked (i) if the defect hinting component is useful, (ii) if the defect hinting component is understandable, (iii) whether it helped to gain new insights into the defect identification process, (iv) it improved the decision-making process, (v) and if the participant would use the defect hinting component in their working environment. For each question, an answer could be provided on a Likert scale between one (strongly disagree) and four (strongly agree). In the second survey, we asked for demographic data, technical background information, and overall feedback regarding the experiments performed. The demographic data we collected were the participants' gender, age range, and education level. Regarding the technical background information, we asked them whether they (a) have basic skills related to manipulation of standard software applications, (b) did some programming in the past but not regularly (on a monthly basis), (c) regularly work on programming/software development, (d) can identify, locate, retrieve, store, organize and analyze digital information and evaluate relevance and purpose, (e) can solve digital problems and explore new ways to take advantage of technology, (f) are aware of basic principles in the area of Machine learning and Artificial Intelligence, and (g) can solve problems using Machine learning and Artificial Intelligence techniques. Finally, we asked them to evaluate the defect hinting components from different experiments, considering whether the particular experiment setups are (a) useful, (b) understandable, (c) informative, (d) help them to be more efficient and (e) learn fast. The last survey question was devoted to understanding if they would adopt a specific setup in their working environment.

We considered two groups of participants: (A) four researchers and students from an artificial intelligence laboratory and (B) three operators tasked with visual quality inspection at the *Philips* manufacturing plant. The goal of having two distinct groups of participants was to understand whether the different backgrounds and domain knowledge influence the task execution and perception of defect hinting. We provide their demographic and technical literacy details in Table 1. All the participants signed consent forms regarding the experiments performed.

Participant	Demographic data			Technical literacy						
	Gender	Age	Education	a	b	c	d	e	f	g
A1	Female	18-30	University Degree							x
A2	Male	18-30	University Degree	x	x	x	x	x	x	x
A3	Male	18-30	Post-graduate degree	x		x	x	x	x	x
A4	Female	31-45	Post-graduate degree			x	x	x	x	x
B1	Female	45-60	Higher education							
B2	Female	45-60	Higher education							
B3	Male	31-45	Higher education				x			

Table 1: Demographic and educational information regarding the participants. Participants' ID is prefixed with the letter of their corresponding group (A for researchers and students of the artificial intelligence laboratory, and B for *Philips* operators.

5 Experiments

In this research, we were interested in exploring means of increasing the velocity and quality of the labeling process. To that end, we performed six experiments, analyzing how the class imbalance and inclusion of a defect hinting component affect the manual revision process. We detail them below:

- **Experiment 1**: stream of images with raw data imbalance without defect hinting. The experiment's goal was to provide a baseline regarding the current manual visual inspection setup.

- **Experiment 2**: balanced stream of images, mitigating the class imbalance with GAN-generated images of defective products (identical images were generated) and no defect hinting.
- **Experiment 3**: stream of images with natural data imbalance and DRAEM anomaly heatmap for defect hinting.
- **Experiment 4**: balanced stream of images (same as Experiment 2) with DRAEM anomaly heatmap for defect hinting.
- **Experiment 5**: balanced stream of images (same as Experiment 2) with a GradCAM heatmap for defect hinting.
- **Experiment 6**: balanced stream of images (same as Experiment 2) where the nearest labeled image (considering the structural similarity index measure) and its label were provided for defect hinting.

Above we describe six experiments. In Experiments 1-4 we aim to determine whether the class imbalance affects the quality of labeling, regardless the existence of some defect hinting. We considered that a higher labeling quality among balanced streams could signal enhanced participants' attention. On the other hand, in Experiments 4-6 we remove the class imbalance and aim to determine which defect hinting technique issues the best results in terms of time and labeling quality. We avoided experiments contrasting balanced and imbalanced settings across all of the defect hinting techniques, given the difficulty to gather participants to collaborate on them, and the additional time required to successfully complete all of them.

The heatmaps are data visualizations that aim to provide an overview of information of interest on a 2D scale, using a color scale to direct the users' attention towards the most interesting regions. In particular, the DRAEM anomaly heatmap highlights the regions that do not conform to what is considered a good sample and therefore can be potentially considered a defect. On the other hand, the GradCAM heatmap plots the values obtained from the gradient of the output of a particular layer, to highlight which image regions are being considered by the model for the task at hand. Ideally, a GradCAM heatmap should highlight image regions provide information regarding whether a defect is present at the image or not.

6 Results and Analysis

The results we obtained from the experiments are twofold. First, the data collected in the web application was analyzed to understand whether differences in labeling quality and times have given different experiment setups. Second, the responses obtained from the surveys completed by the participants were analyzed to get a better understanding of their perception and experience when labeling the data.

6.1 Application measurement results

Experiment	Labeling quality			Labeling time (s)			Unidentified defects	
	Precision↑	Recall↑	F1↑	Median↓	Mean↓	Std. Dev.↓	Other↓	Unable to tell↓
1	0,2857	0,3576	0,3090	*3*	4,27	5,85	36,00	34,33
2	*0,9166*	*0,6304*	*0,7444*	**2**	3,08	5,67	*9,75*	21,00
3	0,6774	0,4983	0,5660	**2**	**2,67**	*5,38*	16,67	11,50
4	0,7339	0,4691	0,5693	**2**	3,00	5,71	10,67	8,00
5	0,8238	0,5212	0,6374	**2**	*2,83*	**5,17**	13,00	*7,75*
6	**0,9393**	**0,8360**	**0,8831**	**2**	3,22	5,82	**1,50**	**5,25**

Table 2: The table summarizes the measurements obtained for the researchers and students of an artificial intelligence laboratory regarding labeling quality and time, along with the number of cases where the user could not tell whether there was a defect or considered the defect did not correspond to one of the existing classes. The best results are bolded, and the second-best results are highlighted in italics.

Experiment	Labeling quality			Labeling time (s)			Unidentified defects	
	Precision↑	Recall↑	F1↑	Median↓	Mean↓	Std. Dev.↓	Other↓	Unable to tell↓
1	0,2468	0,7361	0,3691	**2**	*4,67*	11,34	35,00	**1,00**
2	*0,7915*	*0,8196*	*0,8014*	**2**	5,23	**9,42**	56,33	*2,67*
3	0,4156	0,7845	0,5159	**2**	5,45	12,78	**14,33**	**1,00**
4	0,7103	0,7056	0,7042	*3*	8,67	14,81	109,00	**1,00**
5	0,7721	0,7132	0,7312	**2**	**4,16**	*9,78*	42,33	**1,00**
6	**0,8170**	**0,8598**	**0,8303**	**2**	6,14	13,85	*18,67*	**1,00**

Table 3: The table summarizes the measurements obtained for the *Philips* operators regarding labeling quality and time, along with the number of cases where the user could not tell whether there was a defect or considered the defect did not correspond to one of the existing classes. The best results are bolded, and the second-best results are highlighted in italics.

We summarize the results obtained from the application measurements in Table 2 and Table 3. From the results we observed that the *Philips* operators consistently achieved a high recall, while the recall measured for researchers was low for all of the experiments, except for Experiment 6. We consider the most probable reason for such responses is a different understanding of the quality process. While the researchers are not concerned for the mislabeling a defect, the *Philips* operators avoid a true defect going outside of the manufacturing process. Therefore, when uncertain, they most likely label a manufactured piece as defective and sacrifice good products to avoid leaking defective ones. While the labeling quality differed between both groups, we observed that for both, the best performance was achieved for Experiment 6 and the second-best for Experiment 2 for Precision, Recall, and F1. Given that both experiments correspond to a balanced data stream, we consider the balanced stream increased the participants' attention and that the defect hinting from Experiment 6 was beneficial towards achieving better labeling quality. Experiment 6 also had the lowest or second-lowest number of unidentified defects or cases where the participants could not tell whether a defect was present. Finally, both groups showed the best and

second-best mean labeling speed and low standard deviation in Experiment 5, which had defect hints created with the GradCAM XAI technique. Nevertheless, the median remained the same for most of the experiments, and the quality of labeling was notably lower than the one obtained for Experiment 6.

6.2 Survey results

Exp.	Group	Useful?	Understandable?	New insights?	Improves decision-making?	Would try in working environment?
3	A	**2**	**2**	**2**	**2**	**2**
	B	2	2	1	2	1
4	A	**2**	**2**	**3**	**2**	**2**
	B	1	2	1	1	1
5	A	**3**	**3**	**2**	**2**	**3**
	B	1	2	1	1	2
6	A	**4**	**4**	**4**	**4**	**4**
	B	1	2	2	2	2

Table 4: The table shows the summarized evaluation outcomes obtained from the surveys taken at the end of each experiment (Exp.) for both groups of participants. The results correspond to a Likert scale between one (strongly disagree) and four (strongly agree). With no exception, group A (participants from a laboratory of artificial intelligence) ranked the application higher than group B (*Philips* operators).

Among the survey results, we first analyzed the surveys targeting the participants' perceptions regarding Experiments 3-6 (see Table 4). We found that the perceptions of both groups were considerably different. While the operators mostly responded with values from the lower end of the scale, the researchers responded with higher rates on the Likert scale. The researchers considered the best defect hinting strategy was to provide the image and label of the closest labeled image (Experiment 6). Therefore, the researchers assigned the highest Likert scale value and enforced their decision in the final evaluation survey. They responded that they would prefer the Experiment 6 setup in a working environment. On the other hand, while the highest-ranking experiment among the operators was Experiment 6, their maximum rating was two on the Likert scale of four, tending towards disapproval. Furthermore, when asked which experiment set up they would prefer in a working configuration, they unanimously responded for Experiment 1 or No experiment, which are equivalent responses: to avoid introducing changes to the current working flow.

When comparing DRAEM defect hinting under balanced and imbalanced streams of data and the GradCAM explanations, we observed conflicting responses and a consensus that there is no clear value on such kinds of defect hints for the labelers. Nevertheless, the picture was different when we asked the participants to evaluate all the experiments once all experiments had concluded. The researchers from the artificial intelligence laboratory considered that defect hinting was useful, informative, helped to be more efficient, and learn fast if the defect hints were created based on DRAEM heatmaps or nearest labeled images (see Table 5). The responses provided by the operators were not informative (see Table 6): while a participant considered all of the experiments were informative, another one considered all experiments were understandable. A third participant provided an ambiguous response, which was excluded from the results to avoid interpretive bias.

Exp.	Useful?	Understandable?	Informative?	Helps to be efficient?	Helps to learn fast?
1	0,50	0,50	0,25	0,50	0,25
2	0,25	0,75	0,25	0,50	0,25
3	0,25	0,75	0,50	0,50	0,50
4	0,25	0,75	0,50	0,25	0,25
5	0,25	0,75	0,50	0,25	0,25
6	1,00	1,00	1,00	0,75	1,00

Table 5: The table presents the outcomes of the overall evaluation of the experiments (Exp. 1 through Exp. 6) by the participants from an artificial intelligence laboratory. Responses were provided as a Boolean value, and the table summarizes the proportion of participants answering positively to the questions presented above. The evaluation was based on a survey once all the experiments were completed.

Exp.	Useful?	Understandable?	Informative?	Helps to be efficient?	Helps to learn fast?
1	0,00	0,33	0,33	0,00	0,00
2	0,00	0,33	0,33	0,00	0,00
3	0,00	0,33	0,33	0,00	0,00
4	0,00	0,33	0,33	0,00	0,00
5	0,00	0,33	0,33	0,00	0,00
6	0,00	0,33	0,33	0,00	0,00

Table 6: The table presents the outcomes of the overall evaluation of the experiments (Exp. 1 through Exp. 6) by the participants from *Philips*. Responses were provided as a Boolean value, and the table summarizes the proportion of participants answering positively to the questions presented above. The evaluation was based on a survey once all the experiments were completed.

6.3 Perceptions vs. data

From the results presented in the subsections above, we found that the group of students and researchers in the artificial intelligence laboratory perceived the usefulness of the approach we developed to assist in the manual inspection and that the best-perceived setup matches the best-performing experiment (Experiment 6). However, on the other side, the experiments were not perceived as helpful by the *Philips* operators. Nevertheless, the data gathered from the application shows that while the median labeling time remained the same as in their current setup, and the mean labeling time was 30% higher, the quality of labeling was notably superior, resulting in more than three times greater precision, and more than two times the original F1 score. Furthermore, Experiment 6 helped the *Philips* operators reduce the number of unidentified defects by more than 80%. Therefore, we conclude that the setup used for Experiment 6 has shown benefits regardless of the participants' perceptions. Furthermore, further work needs to be done to determine how such changes could be communicated to the operators to change their perceptions.

7 Conclusion

This research presents a set of experiments designed to understand which machine learning technologies and approaches could be leveraged to enhance the manual revision process for those products whose quality cannot be determined with certainty with a machine learning model. To that end, we conducted six experiments with six hundred images each, based on a real-world dataset provided by *Philips Consumer Lifestyle BV* and evaluated two different groups of participants: (a) researchers and students from an artificial intelligence laboratory, and (b) *Philips* operators tasked with the visual inspection of products. We then compared data collected from the application during the experiments' execution with surveys. We analyzed which experiments yielded the best performance and whether the perception of each group matched the objective measures. We found the best performance was achieved with a balanced stream of images (an equal amount of images considering good and defective products) and hinting the user on the expected class by showing the most similar labeled image and the corresponding class. Gains in labeling precision were at least three times higher than in the existing setup, while the median labeling time remained the same or slightly decreased. Furthermore, the number of unidentified defects was reduced by more than 80% in the worst case. Future work will explore new setups that could enhance the manual revision process by introducing slight imbalances that could reduce the synthetic images overhead and monitor the operators' wellbeing, ensuring tasks are changed, or different strategies adopted when their attention decreases.

Acknowledgments

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 program project STAR under grant agreement number H2020-956573.

The authors acknowledge the valuable input and help of Paulien Dam and Yvo van Vegten from *Philips Consumer Lifestyle BV*.

References

Beltran-Gonzalez, C., Bustreo, M., and Del Bue, A. (2020). External and internal quality inspection of aerospace components. In *2020 IEEE 7th International Workshop on Metrology for AeroSpace (MetroAeroSpace)*, pages 351–355. IEEE.

Benbarrad, T., Salhaoui, M., Kenitar, S. B., and Arioua, M. (2021). Intelligent machine vision model for defective product inspection based on machine learning. *Journal of Sensor and Actuator Networks*, 10(1):7.

Carvajal Soto, J., Tavakolizadeh, F., and Gyulai, D. (2019). An online machine learning framework for early detection of product failures in an industry 4.0 context. *International Journal of Computer Integrated Manufacturing*, 32(4-5):452–465.

Chouchene, A., Carvalho, A., Lima, T. M., Charrua-Santos, F., Osorio, G. J., and Barhoumi, W. (2020). Artificial intelligence for product quality inspection toward smart industries: quality control of vehicle non-conformities. In *2020 9th international conference on industrial technology and management (ICITM)*, pages 127–131. IEEE.

Defard, T., Setkov, A., Loesch, A., and Audigier, R. (2021). Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer.

Frank, A. G., Dalenogare, L. S., and Ayala, N. F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics*, 210:15–26.

Gobert, C., Reutzel, E. W., Petrich, J., Nassar, A. R., and Phoha, S. (2018). Application of supervised machine learning for defect detection during metallic powder bed fusion additive manufacturing using high resolution imaging. *Additive Manufacturing*, 21:517–528.

Iglesias, C., Martinez, J., and Taboada, J. (2018). Automated vision system for quality inspection of slate slabs. *Computers in Industry*, 99:119–129.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., and Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36.

Samek, W. and Müller, K.-R. (2019). Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22. Springer.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Wuest, T., Irgens, C., and Thoben, K.-D. (2014). An approach to monitoring quality in manufacturing using supervised machine learning on product state data. *Journal of Intelligent Manufacturing*, 25(5):1167–1180.

Yang, J., Li, S., Wang, Z., Dong, H., Wang, J., and Tang, S. (2020). Using deep learning to detect

defects in manufacturing: a comprehensive survey and current challenges. *Materials*, 13(24):5755.

Zavrtanik, V., Kristan, M., and Skočaj, D. (2021). Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339.