Explainable Artificial Intelligence Meets Active Learning: A Novel GradCAM-based Active Learning Strategy.

Karel Križnar[*]
University of Ljubljana
Faculty of mathematics and physics
Jadranska ulica 19, Ljubljana, Slovenija
kk4876@student.uni-lj.si

Jože M. Rožanec[†]
Jožef Stefan International Postgraduate School
Jamova cesta 39, 1000 Ljubljana, Slovenia
joze.rozanec@ijs.si

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

Dunja Mladenić
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
dunja.mladenic@ijs.si

Abstract. *Active learning is a subfield of machine learning that studies how to identify data instances that contribute most to the learning of a given learner and requests some oracle to provide complementary information to enhance the learner's learning process towards a specific goal. This paper proposes a novel active learning strategy for image classification tasks. Furthermore, we obtain insights that lay the ground for developing a novel early-stopping criteria to maximize the model's learning while reducing the required instances. The proposed active learning technique uses insights obtained via GradCAM activation maps to understand the cognition process triggered by each image in the learner and select those that trigger the most diverse patterns, presupposing that diverse GradCAM patterns point to new learning opportunities. To our knowledge, the approach is among the first to use insights from explainable artificial intelligence to drive data selection and annotation. Nevertheless, our results show that it does not achieve the same quality of results as the random and uncertainty sampling techniques we compared to. More research is required to understand how to enhance the models' performance with such a strategy.*

Keywords. Active Learning; Automated Visual Inspection; Early Stopping; Machine Learning; Explainable Artificial Intelligence; Neural Networks; Quality Assurance and Maintenance

1 Introduction

The Industry 4.0 paradigm fosters using advanced technologies to optimize manufacturing further by increasing the flexibility and efficiency of the manufacturing process and the product value over the whole manufacturing lifecycle (Frank et al., 2019). Among the enabling technologies, we find (i) the Internet of Things, (ii) Big Data, (iii) Cloud Computing, (iv) Digital Twin, (v) Additive Manufacturing, (vi) Autonomous Robots, (vii) Simulation, (viii), Cybersecurity, (ix) Augmented Reality, and (x) Artificial Intelligence (Suleiman et al., 2022). These technologies are used to improve manufacturing sustainability in multiple aspects, such as employee productivity, improved profit margin, intelligent production planning and control, manufacturing agility, and efficiency, among others (Ching et al., 2022).

Product quality is relevant for businesses as it fosters customer trust, enhances loyalty, and strengthens the brand's reputation (Yang et al., 2020). Therefore, quality control plays a key role in manufacturing, given it ensures compliance with standards and specifications. During the visual inspection, integrity, surface finish, and geometric dimensions can be assessed (Newman and Jain, 1995). The automated visual inspection aims to mitigate common issues that relate to the subjective nature of human inspection (e.g., operator-to-operator inconsistency and quality dependence on the employees' experience and well-being (See, 2012)), and scalability issues (e.g., speed of inspection and continuous execution of the inspection process among others (Chouchene et al., 2020)).

Machine vision is a sub-field of artificial intelligence that develops means to process images with different purposes. It is used in the context of visual inspection to detect and flag faulty products while recognizing their defects and allowing rapid intervention to remove such products, address failure root causes, and mitigate the void that results from the estimated and actually delivered manufactured products in the production pipeline (Javaid et al., 2022). The decreased cost of sensors, the use of machine learning algorithms,

[*] Karel Križnar and Jože M. Rožanec are co-first authors with equal contribution and importance

[†] Corresponding author: Jože M. Rožanec. Email: joze.rozanec@ijs.si

and other technologies are key enablers for the broader adoption of automated visual inspection in the context of Industry 4.0. Furthermore, an automated visual inspection can be considered an essential element of manufacturing and can have a broader impact on the overall organization under the Industry 4.0 paradigm (Konstantinidis et al., 2021).

Supervised machine learning models require labeled data to learn specific defect types. Labeling data is expensive. Furthermore, machine learning model training times have associated costs that must be considered to determine the value for money (Justus et al., 2018). Reducing training times has a direct impact on reducing such costs. Active learning provides means to select data instances to speed up the learning of a machine learning algorithm and therefore reduce training times, given fewer data instances are required to achieve certain discriminative performance.

The main contribution of this research is the development of a novel active learning technique and early stopping criteria. While the active learning criteria aims to achieve a steep learning curve for the machine learning algorithm, the early stopping criteria allows to identify when training should be stopped, given that the algorithm will not learn much more. Combining them, we aim to achieve an enhanced learning process of the machine learning algorithms while reducing the amount of labeled data that must be fed to the algorithm and the training times required to train a model. The experiments were performed considering two real-world use cases: a dataset provided by *Philips Consumer Lifestyle BV* corporation and an open dataset from Kolektor (Božič et al., 2021), which has become one of the standard datasets for automated visual inspection tasks.

To evaluate machine learning models, we measure the discriminative capability with the AUC ROC metric. Furthermore, we compare how many data instances are required by each active learning strategy to achieve their best performance and how many samples must be shown to the model before early stopping.

The rest of this paper is structured as follows: Section 2 presents related work, Section 3 introduces a novel active learning strategy, while the Section 4 describes the *Kolektor* and *Philips Consumer Lifestyle BV* use cases and datasets. Section 5 introduces the methodology and experiments. Section 6 describes and analyzes the results obtained. Finally, Section 7 offers our conclusions and outlines future work.

2 Related Work

Since deep learning algorithms for machine vision have achieved super-human performance on certain tasks (Ciregan et al., 2012), much research has been invested in how to enhance them. Machine learning models for automated visual inspection have been widely adopted, and state-of-the-art performance has been achieved with deep learning models (Pouyanfar et al., 2018; Beltrán-González et al., 2020). Nevertheless, training such models from scratch is computationally expensive. Furthermore, acquiring enough quality labeled data for training such models requires a considerable effort (Kocaguneli et al., 2012; Whang et al., 2023). Therefore, approaches such as transfer learning or few-shot learning have been proposed to reduce both (Zhuang et al., 2020; Wang et al., 2020). While both approaches exist independently of active learning, active learning can be used to enhance both.

While active learning has been researched in the past, attention to it has been gradually increasing as a means to reduce the labeling effort (Ren et al., 2021). To perform data selection, active learning methods mostly considered either the inherent characteristics of the data (e.g., informativeness, representativeness, and diversity (Wu, 2018)) or the model's behavior w.r.t. the data (e.g., the closeness of a datapoint to the decision boundary). Such methods considered the data selection process dissociated from the actual model's learning process. Nevertheless, there have been some exceptions (e.g., (Zhu et al., 2019), who proposed optimizing both objectives at once by formulating data selection as a robust optimization problem). We consider that active learning data selection policies developed up to now used inherent data properties or model behavior as proxies to understanding the actual model's learning process and decision boundaries evolution. The advent of explainable artificial intelligence has opened a wide range of opportunities. Explainable artificial intelligence is a sub-field of artificial intelligence concerned with enabling humans to understand machine learning models, trust and effectively manage them (Arrieta et al., 2020). While much effort is being invested into developing new explainable models, explainability techniques, and metrics to assess their quality, to our knowledge, insights obtained from explainable artificial intelligence have not been leveraged yet to drive data selection in an active learning context. Nevertheless, such insights enable selecting data with a more direct understanding of the ongoing learning process inside the machine learning model.

Multiple explainability techniques have been devised to provide insights into machine vision models. Among the most popular techniques we find GradCAM (Selvaraju et al., 2017) (it uses the gradient information to understand how strongly the neurons activate), DeepLIFT (Shrikumar et al., 2016) (uses a derivative-based method to propagate activation differences, and determine how changes in the image would affect predictions), or Smooth-Grad (Smilkov et al., 2017) (measures local sensitivity based on small image perturbations). Many ways have been devised to convey explainable artificial intelligence outcomes for machine vision. Authors have considered overlaying the original image with a cloud of points, an activation or heat map, or some outline (Hudon et al., 2021), displaying only a relevant part of the image (Ribeiro et al., 2016), or highlighting it (Buhrmester et al., 2021).

While Meng et al. (Meng et al., 2020) consider there is a research void regarding using active learning manufacturing domain, we have found some examples where it was successfully applied to the quality inspection use cases. In particular, Van et al. (van Garderen, 2018) explored how active learning could enhance the learning process of a machine learning model measuring the local displacement between layers on a chip. In the same line, Shim et al. (Shim et al., 2020) described a technique developed to select wafer maps, which provide key information to engineers for detecting root causes of failure in the semiconductor manufacturing process. Finally, Dai et al. (Dai et al., 2018) described how active learning was used to automatically enlarge a dataset when training a model to recognize solder joint defects in printed circuit boards.

3 GradCAM$_{\text{avg}}$

So far, Active learning techniques have mainly focused on the characteristics of the data or the models' behavior to find instances that would be helpful to refine the decision boundary. Nevertheless, we consider both approaches as proxies to direct insights into models' rationale, which can now be obtained using explainable artificial intelligence techniques. One such explainable artificial intelligence technique is GradCAM (Selvaraju et al., 2017), which computes the importance map by taking the derivative of the reduction layer output for a given class with respect to a convolutional feature map. The importance map can be rendered as an activation map for visualization purposes.

For this research, we have developed a novel active learning technique named GradCAM$_{\text{avg}}$, which consists of generating GradCAM activation maps for all of the images in the pool and computing the structural similarity index measure (SSIM) between the average GradCAM representation of the images seen so far and the rest of the unlabeled images. The unlabeled images are then selected considering those most dissimilar compared to the average GradCAM image. The intuition behind the algorithm is that explainable artificial intelligence techniques can provide a means to understand how the model perceives each unlabeled image and the ones seen so far. The unlabeled images whose GradCAM activation map is most dissimilar from those seen so far are likely to offer more novel information to the model and result in a steeper learning curve. We detail the procedure in Algorithm 1.

Little work has been done combining explainable artificial intelligence and active learning. E.g., Ghai et al. (2021) used explainable artificial intelligence insights to reduce opaqueness regarding the samples presented to the human oracle labeling the data. To our knowledge, GradCAM$_{\text{avg}}$ is among the first active learning techniques leveraging insights from explainable artificial intelligence to boost models' learning. Concurrently similar approaches have been developed by Zajec et al. (2023). We do not know of other methods leveraging explainable artificial intelligence in the context of active learning.

4 Use Cases

The research we performed was based on two real-world use cases concerning the visual inspection of products manufactured by two European companies (*Kolektor Group d.o.o.* from Slovenia and *Philips Consumer Lifestyle BV* from The Netherlands).

Kolektor Group d.o.o. published a dataset (Kolektor SDD2 a.k.a. KSDD2) of over 3.000 images based on a real-world example (Božič et al., 2021). The dataset was constructed from color images captured in a controlled environment by a visual inspection system and annotated by the company and a group of researchers. The dataset has been considered in several publications related to automated visual inspection. We adopt it as a reference dataset for which the results obtained can be compared later with further research on this topic.

Philips Consumer Lifestyle BV has printing machine setups for several products, which are manually handled and inspected to determine their quality. If a defect is observed, the manufactured product is removed from the manufacturing line to ensure only products conforming to high-quality standards are left. Human inspectors spend several seconds handling, inspecting, and labeling the products. If automating the visual inspection process, it is expected that human inspectors would require several seconds to label each product to create an initial dataset. Therefore, any gains in reducing the needed labeled data and labeling times directly translate into cost savings. *Philips Consumer Lifestyle BV* provided a dataset of over 3,000 annotated images. While fine-grained labels were supplied, we trained the models to recognize only whether a defect exists in the manufactured product.

5 Methodology and Experiments

5.1 Methodology

When training the machine learning models, we performed stratified cross-validation (Zeng and Martinez, 2000). We adopted *k=10* based on recommendations by Kuhn et al. (2013) and considered one fold for testing, one for validation, and the rest was used to simulate a pool of unlabeled data for active learning, from which we sourced the data samples. We used a ResNet-18 network (He et al., 2016) and trained it incrementally by sourcing batches of 32 images from the active learning pool until pool exhaustion.

Algorithm 1 $GradCAM_{avg}$ active learning algorithm. We select batches of images based on how different their GradCAM activation maps are from the average GradCAM activation map obtained from the images labeled so far.

```
$Set_{UnlabeledImages} \leftarrow \{img_i, img_{i+1}, img_{i+2}, \ldots, img_n\}$
$Set_{LabeledImages} \leftarrow \{img_j, img_{j+1}, img_{j+2}, \ldots, img_m\}$
procedure SELECTINSTANCES($Set_{UnlabeledImages}$, $Set_{LabeledImages}$, $n_{images\,to\,select}$)
    $Set_{GradCAM_{LabeledImages}} \leftarrow \emptyset$
    $Set_{GradCAM_{Unlabeled}} \leftarrow \emptyset$
    for $img \in Set_{LabeledImages}$ do $Set_{GradCAM_{LabeledImages}} \cup \{GradCAM(img)\}$
    end for
    for $img \in Set_{UnlabeledImages}$ do $Set_{GradCAM_{Unlabeled}} \cup \{GradCAM(img)\}$
    end for
    $GradCAM_{avg} \leftarrow avg(Set_{GradCAM_{LabeledImages}})$
    $Dictionary_{GradCAM_{Unlabeled}\,to\,SSIM} \leftarrow \{\}$
    for $GradCAM_{img} \in Set_{GradCAM_{Unlabeled}}$ do
        $Dictionary_{GradCAM_{Unlabeled}\,to\,SSIM} \leftarrow GradCAM_{img}, d_{SSIM}(GradCAM_{avg}, GradCAM_{img})$
    end for
     return $top(Dictionary_{GradCAM_{Unlabeled}\,to\,SSIM}, n_{images\,to\,select})$
end procedure
```

5.2 Experiments

To understand the inherent performance limitations of the machine learning models on the datasets, we trained a model on all of the data available in the active learning pool. We considered the resulting performance to be the best that could be obtained by a ResNet-18 model trained on that particular data.

We trained the ResNet-18 model considering three active learning strategies: (i) random sampling (baseline), (ii) uncertainty sampling (selects samples with the highest uncertainty, computed as the difference between the highest predicted score and one), and (iii) a novel technique of our own that we named GradCAM$_{avg}$, which we described in Section 3. We examined the models' performance over time for each case and analyzed when the learning performance peaked. Furthermore, we took the subset of images presented to the model until learning peaked in the active incremental learning and trained the model with it in a batch setting. This enabled us to understand differences in performance between models trained in batch or incremental learning settings while reducing the amount of data shown to them based on the active learning criteria.

6 Results and Analysis

6.1 Active learning: batch vs. incremental learning

Among the experiments, we were first interested in assessing how data sampling strategies affected the learning process and discriminative performance of the ResNet-18 model. Regarding discriminative performance (see Table 1), we found that the random sampling led to the highest AUC ROC score (0.8967) for the Kolektor SDD2 dataset, while uncertainty sampling and GradCAM$_{avg}$ lagged with an AUC ROC score of 0.8647 and 0.8502 in an incremental learning setting. Nevertheless, when considering the same sampled images in a batch learning setting, the differences were less pronounced: random and uncertainty sampling achieved almost the same mean AUC ROC (0.9103 and 0.9187), and GradCAM$_{avg}$ slightly lagged (0.9006). For the Philips shavers dataset, we verified the same performance pattern concerning active learning techniques in the incremental learning setting. Nevertheless, uncertainty sampling and GradCAM$_{avg}$ lagged behind random sampling when training the model in a batch setting.

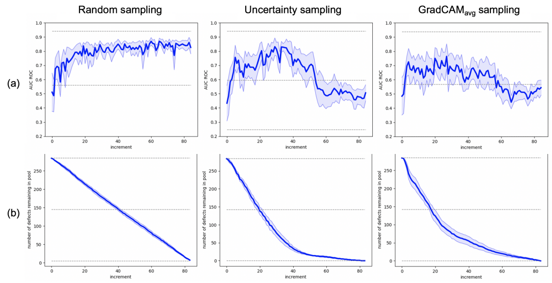

Figure 1: We display plots in two rows: (a) discriminative performance over time and (b) number of images regarding defective products left in the active learning pool over time. The plots correspond to the Kolektor SDD2 dataset.

6.2 Active learning: learning curves vs. pool composition

When analyzing the results (see Fig. 1 and Fig. 2), we focused on two plots: (a) AUC ROC plots, display-

AUC ROC (average)		**Active Learning strategy (data selection)**		
Dataset	**Learning strategy**	**Random**	**Uncertainty**	**GradCAMavg**
Kolektor SDD2	**Incremental Learning**	0.8967±0.0133	0.8647±0.0098	0.8502±0.0310
	Batch Learning	0.9103±0.0161	0.9187±0.0215	0.9006±0.0302
Philips (shavers)	**Incremental Learning**	0.9416±0.0100	0.9124±0.0123	0.9095±0.0162
	Batch Learning	0.9799±0.0088	0.9526±0.0086	0.9395±0.0113

Table 1: AUC ROC for binary classification setting, comparing three active learning techniques and how the same set of images affects learning when considering incremental or batch learning.

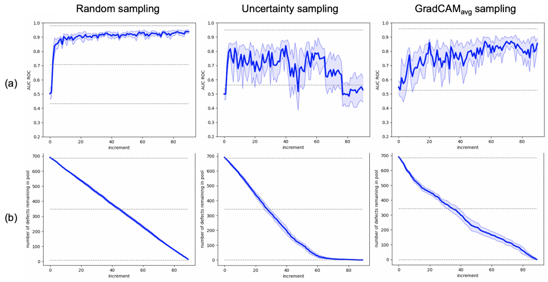

Figure 2: We display plots in two rows: (a) discriminative performance over time and (b) number of images regarding defective products left in the active learning pool over time. The plots correspond to the Philips Consumer Lifestyle BV shavers dataset.

ing how models' discriminative performance evolved and (b) plots showing the number of samples of defective products left in the pool of unlabeled data. We found that the AUC ROC curve usually started decreasing close to the knee point of the plot (b).

The curve in plot (b) strongly reminded the scree plot curves. Based on this observation, we wondered whether a similar rule of thumb as the one used to select principal components in principal component analysis with a scree plot (Satopaa et al., 2011) could be used in this case to determine how long to sample data from the active learning pool. While the information displayed in plot (b) assumes the labels are known, the active learning setting assumes unlabeled data. Nevertheless, some soft labeling strategies could be used to mitigate this issue.

When considering the results presented in Table 2 and Table 3, we found that in all cases, the model with the best discriminative performance in the validation set until the knee point in the plot had better performance than the model at the knee point. Nevertheless, these models displayed inferior performance in all but one case than the best discriminative model obtained by annotating the data samples until pool exhaustion. Furthermore, the best models lead by at least 0.1 AUC ROC points when compared against the best models obtained until the knee point.

Analyzing the Kolektor SDD2 dataset, we found that the best-performing models originated after the knee point for random sampling. This was not the case for the uncertainty sampling and GradCAM$_{avg}$ strategies where, on average, the best models appeared before the knee point, but not always. On the other hand, in the Philips shavers dataset, the best models were observed after the knee point when using random and GradCAM$_{avg}$ sampling. This was not the case for uncertainty sampling, where, on average, the best models appeared before the knee point. Moreover, the pattern of having the best models appear before the knee point seems to correlate to the shape of the plot (b): the best models appear before the knee point when the plots have a part of the curve that resembles debris fallen from a mountain and lying at its base. While uncertainty sampling displayed a consistent behavior across both datasets, early stopping at the knee point significantly degraded the model's performance when trained incrementally.

7 Conclusion

In this research, we describe a novel active learning technique we named GradCAM$_{avg}$ given it aims to select candidate images for labeling from unlabeled data based on their SSIM score when compared against the average of GradCAM activation maps for the images labeled up to that point in time. To our understanding, this method is among the first ones leveraging insights from explainable artificial intelligence techniques to drive data selection. When analyzing AUC ROC plots and plots describing the composition of the unlabeled dataset, we found that active learning strategies that greedily source data instances of defective products tend to achieve best-performing models before the knee point at those plots. While applying such criteria led to sub-optimal results in this research, more effort is required to overcome this limitation and capitalize on this insight to develop a novel early stopping criteria. We envision that such an early stopping criteria could guide active learning data annotation efforts. Future work will attempt to evaluate the GradCAM$_{avg}$ technique on a comprehensive number of datasets and develop a novel early-stopping technique based on the insights of this research.

Number of samples seen by the model		Active Learning strategy (data selection)		
Dataset	**Cut-off criteria**	**Random**	**Uncertainty**	**GradCAMavg**
Kolektor SDD2	**Best Discriminative Performance**	1502±416	1088±160	672±288
	Knee Point in Plot	1184±256	1184±96	800±160
	Best Discriminative Performance until Knee Point	896±320	1024±128	448±192
Philips (shavers)	**Best Discriminative Performance**	2272±384	960±352	2112±416
	Knee Point in Plot	1600±352	1664±128	768±256
	Best Discriminative Performance until Knee Point	1792±384	768±320	544±224

Table 2: Number of samples for binary classification setting, comparing three active learning techniques. We aim to understand whether the best models are learned before or after the knee point. Furthermore, we are interested in understanding when the best model before a knee point originates.

AUC ROC (average)		Active Learning strategy (data selection)		
Dataset	**Cut-off criteria**	**Random**	**Uncertainty**	**GradCAMavg**
Kolektor SDD2	**Best Discriminative Performance**	0.8967±0.0133	0.8647±0.0098	0.8502±0.0310
	Knee Point in Plot	0.7965±0.0348	0.7183±0.0665	0.6646±0.1236
	Best Discriminative Performance until Knee Point	0.8216±0.0298	0.8255±0.0317	0.7137±0.0917
Philips (shavers)	**Best Discriminative Performance**	0.9103±0.0161	0.9187±0.0215	0.9006±0.0302
	Knee Point in Plot	0.9147±0.0195	0.5819±0.1481	0.6445±0.1205
	Best Discriminative Performance until Knee Point	0.9130±0.0182	0.7458±0.0979	0.6942±0.0641

Table 3: AUC ROC for binary classification setting, comparing three active learning techniques. We aim to understand the performance of the best overall model and how it compares against the best model obtained until the knee point.

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.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Beltrán-González, 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.

Božič, J., Tabernik, D., and Skočaj, D. (2021). Mixed supervision for surface-defect detection: from weakly to fully supervised learning. *Computers in Industry*.

Buhrmester, V., Münch, D., and Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, 3(4):966–989.

Ching, N. T., Ghobakhloo, M., Iranmanesh, M., Maroufkhani, P., and Asadi, S. (2022). Industry 4.0 applications for sustainable manufacturing: A systematic literature review and a roadmap to sustainable development. *Journal of Cleaner Production*, 334:130133.

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.

Ciregan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE.

Dai, W., Mujeeb, A., Erdt, M., and Sourin, A. (2018). Towards automatic optical inspection of soldering defects. In *2018 International Conference on Cyberworlds (CW)*, pages 375–382. IEEE.

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.

Ghai, B., Liao, Q. V., Zhang, Y., Bellamy, R., and Mueller, K. (2021). Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–28.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., and Sénécal, S. (2021). Explainable artificial intelligence (xai): how the visualization of ai predictions affects user cognitive load and confidence. In *Information Systems and Neuroscience: NeuroIS Retreat 2021*, pages 237–246. Springer.

Javaid, M., Haleem, A., Singh, R. P., Rab, S., and Suman, R. (2022). Exploring impact and features of machine vision for progressive industry 4.0 culture. *Sensors International*, 3:100132.

Justus, D., Brennan, J., Bonner, S., and McGough, A. S. (2018). Predicting the computational cost of deep learning models. In *2018 IEEE international conference on big data (Big Data)*, pages 3873–3882. IEEE.

Kocaguneli, E., Menzies, T., Keung, J., Cok, D., and Madachy, R. (2012). Active learning and effort estimation: Finding the essential content of software effort estimation data. *IEEE Transactions on software engineering*, 39(8):1040–1053.

Konstantinidis, F. K., Mouroutsos, S. G., and Gasteratos, A. (2021). The role of machine vision in industry 4.0: an automotive manufacturing perspective. In *2021 IEEE international conference on imaging systems and techniques (IST)*, pages 1–6. IEEE.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Meng, L., McWilliams, B., Jarosinski, W., Park, H.-Y., Jung, Y.-G., Lee, J., and Zhang, J. (2020). Machine learning in additive manufacturing: A review. *Jom*, 72(6):2363–2377.

Newman, T. S. and Jain, A. K. (1995). A survey of automated visual inspection. *Computer vision and image understanding*, 61(2):231–262.

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.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

See, J. E. (2012). Visual inspection: a review of the literature. *Sandia Report SAND2012-8590, Sandia National Laboratories, Albuquerque, New Mexico*.

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.

Shim, J., Kang, S., and Cho, S. (2020). Active learning of convolutional neural network for cost-effective wafer map pattern classification. *IEEE Transactions on Semiconductor Manufacturing*, 33(2):258–266.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Suleiman, Z., Shaikholla, S., Dikhanbayeva, D., Shehab, E., and Turkyilmaz, A. (2022). Industry 4.0: Clustering of concepts and characteristics. *Cogent Engineering*, 9(1):2034264.

van Garderen, K. (2018). Active learning for overlay prediction in semi-conductor manufacturing.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Whang, S. E., Roh, Y., Song, H., and Lee, J.-G. (2023). Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, pages 1–23.

Wu, D. (2018). Pool-based sequential active learning for regression. *IEEE transactions on neural networks and learning systems*, 30(5):1348–1359.

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.

Zajec, P., Rožanec, J. M., Theodoropoulos, S., Fontul, M., Koehorst, E., Fortuna, B., and Mladenić, D. (2023). Few-shot learning for defect detection in manufacturing. submitted.

Zeng, X. and Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12.

Zhu, D., Li, Z., Wang, X., Gong, B., and Yang, T. (2019). A robust zero-sum game framework for pool-based active learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 517–526. PMLR.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.