Active Learning for Automated Visual Inspection of Manufactured Products

Elena Trajkova*
University of Ljubljana, Faculty of Electrical Engineering
Ljubljana, Slovenia
trajkova.elena.00@gmail.com

Jože M. Rožanec*
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
joze.rozanec@ijs.si

Paulien Dam
Philips Consumer Lifestyle BV
Drachten, The Netherlands
paulien.dam@philips.com

Blaž Fortuna
Qlector d.o.o.
Ljubljana, Slovenia
blaz.fortuna@qlector.com

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

Quality control is a key activity performed by manufacturing enterprises to ensure products meet quality standards and avoid potential damage to the brand's reputation. The decreased cost of sensors and connectivity enabled an increasing digitalization of manufacturing. In addition, artificial intelligence enables higher degrees of automation, reducing overall costs and time required for defect inspection. In this research, we compare three active learning approaches and five machine learning algorithms applied to visual defect inspection with real-world data provided by *Philips Consumer Lifestyle BV*. Our results show that active learning reduces the data labeling effort without detriment to the models' performance.

CCS CONCEPTS

• **Information systems → Data mining**; • **Computing methodologies → Computer vision problems**; • **Applied computing**;

KEYWORDS

Smart Manufacturing, Machine Learning, Automated Visual Inspection, Defect Detection

ACM Reference Format:
Elena Trajkova, Jože M. Rožanec, Paulien Dam, Blaž Fortuna, and Dunja Mladenić. 2021. Active Learning for Automated Visual Inspection of Manufactured Products. In *Ljubljana '21: Slovenian KDD Conference on Data Mining and Data Warehouses, October, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Quality control is one of the critical activities that must be performed by manufacturing enterprises [27, 28]. The main purpose of such activity is to detect product defects meeting quality standards, avoid rework, supply chain disruptions, and avoid potential damage to the brand's reputation [3, 27]. Along with the information regarding defective products, it provides insights into when and where such defects occur, which can be used to further dig into the root causes of such defects and mitigation actions to improve the quality of manufacturing products and processes.

The decreased cost of sensors and connectivity enabled an increasing digitalization of manufacturing [3], which along with the adoption of Artificial Intelligence (AI) [12], represents an opportunity towards enhancing the defect detection in industrial settings [5]. While the quality of the manual inspection has low scalability (requires time to train an inspector, the employees can work a limited amount of time and are subject to fatigue, and the inspection itself is slow), its quality can be affected by the operator-to-operator inconsistency, and it depends on the complexity of the task, the employees (e.g., their intelligence, experience, well-being), the environment (e.g., noise and temperature), the management support and communication [23]; none of these factors affect the outcome of automated quality inspection. Machine learning has been successfully applied to defect detection in a wide range of scenarios [1, 9, 11, 15, 21].

An annotated dataset must be acquired to implement machine learning models for defect detection successfully. The increasing number of sensors provides large amounts of data. As the manufacturing process quality increases, the data obtained from the sensors is expected to be highly imbalanced: most of the data instances will correspond to non-defective products, and a small proportion of them will correspond to different kinds of defects. Annotating all the data is prone to similar limitations as manual inspection described in the paragraph above. It is thus imperative to provide strategies to select a limited subset of them that are most informative to the defect detection models.

We frame the defect detection problem as a supervised learning problem. Given a large amount of unlabeled data, and based on the premise that only a tiny fraction of the data provides new information to the model and thus has the potential to enhance its performance, we adopt an active learning approach. Active learning is a subfield of machine learning that attempts to identify the most informative unlabeled data instances, for which labels are requested some *oracle* (e.g., a human expert) [24]. This research compares three active learning strategies: pool-based sampling, stream-based sampling, and query by committee.

*Both authors contributed equally to this research.

SiKDD '21, October, 2021, Ljubljana, Slovenia

The main contributions of this research are (i) a comparative study between the five most frequently cited machine learning algorithms for automated defect detection and (ii) three active learning approaches (iii) for a real-world multiclass classification problem. We develop the machine learning models with images provided by the *Philips Consumer Lifestyle BV* corporation. The dataset comprises shaver images divided into three classes, based on the defects related to the printing of the logo of the *Philips Consumer Lifestyle BV* corporation: good shavers, shavers with double printing, and shavers with interrupted printing.

We evaluate the models using the area under the receiver operating characteristic curve (AUC ROC, see [4]). AUC ROC is widely adopted as a classification metric, having many desirable properties such as being threshold independent and invariant to a priori class probabilities. We measure AUC ROC considering prediction scores cut at a threshold of 0.5.

This paper is organized as follows. Section 2 outlines the current state of the art and related works, Section 3 describes the use case, and Section 4 provides a detailed description of the methodology and experiments. Finally, section 5 outlines the results obtained, while Section 6 concludes and describes future work.

2 RELATED WORK

Among the many techniques used for automated defect inspection, we find the automated visual inspection, which refers to image processing techniques for quality control, usually applied in the production line of manufacturing industries [1]. Visual inspection requires extracting features from the images, which are used to train the machine learning model. This procedure is simplified when using deep learning models, enabling end-to-end learning, where a single architecture can perform feature extraction and classification [10, 18], and have shown state-of-the-art performance for image classification [20].

The use of automated visual inspection for defect detection has been applied to multiple manufacturing use cases. [21] manually extracted features (e.g., histograms) from machine component images and compared the performance of the Näive Bayes and C4.5 models. [9] extracted statistical features from the images and compared the performance of Support Vector Machines (SVM), Multilayer Perceptron (MLP), and k-nearest neighbors (kNN) models for visual inspection of microdrill bits in printed circuit board production. [11] used 3D convolutional filters applied on computed tomography images and an SVM classifier for defect detection during metallic powder bed fusion in additive manufacturing. [15] used some heuristics to detect regions of interest on slate slab images, on which they performed feature engineering to later train an SVM model on them. Finally, [1] reported using a custom neural network for feature extraction and an SVM model for classification when inspecting aerospace components.

While the authors cited above worked with fully labeled datasets, a production line continually generates new data, exceeding the labeling capacity. A possible solution to this issue is the use of active learning, where the active learner identifies informative unlabeled instances and requests labels to some *oracle*. Typical scenarios involve (i) membership query synthesis (a synthetic data instance is generated), (ii) stream-based selective sampling (the unlabeled instances are drawn one at a time, and a decision is made whether a label is requested, or the sample is discarded), and (iii) pool-based selective sampling (queries samples from a pool of unlabeled data). Among the frequently used querying strategies, we find (i) uncertainty sampling (select an unlabeled sample with the highest uncertainty, given a certain metric or machine-learning model[17]), or (ii) query-by-committee (retrieve the unlabeled sample with the highest disagreement between a set of forecasting models (*committee*)) [6, 24]. More recently, new scenarios have been proposed leveraging reinforcement learning, where an agent learns to select images based on the similarity relationship between the instances and rewards obtained based on the oracle's feedback [22]. In addition, it has been demonstrated that ensemble-based active learning can effectively counteract class imbalance through new labeled images acquisition [2].

Active learning was successfully applied in the manufacturing domain, but scientific literature remains scarce on this domain [19]. Some use cases include the automatic optical inspection of printed circuit boards[8] and the identification of the local displacement between two layers on a chip in the semi-conductor industry[25].

The use of machine learning automates the defect detection, and active learning enables an *inspection by exception* [5], only querying for labels of the images that the model is most uncertain about. While this considerably reduces the volume of required inspections, it is also essential to consider that it can produce an incomplete ground truth by missing the annotations of defective parts classified as false negatives and not queried by the active learning strategy [7].

3 USE CASE

The use case provided for this research corresponds to visual inspection of shavers produced by *Philips Consumer Lifestyle BV*. The visual quality inspection aims to detect defective printing of a logo on the shavers. This use case focuses on four pad printing machines setup for a range of different products, and different logos. A lot of products are produced every day on these machines, which are manually handled and inspected on their visual quality and removed from further processing if the prints on the products are not classified as good. Operators spend several seconds handling, inspecting, and labeling the products. Given an automated visual quality inspection system would strongly reduce the need to manually inspect and label the images, it could speed up the process for more than 40%. Currently there are two types of defects classified related to the printing quality of the logo on the shaver: double printing, and interrupted printing. Therefore, images are classified into three classes: good printing (class zero), double printing (class one), and interrupted printing (class two). A labeled dataset with a total of 3.518 images was provided to train and test the models.

4 METHODOLOGY

We pose automated defect detection as a multiclass classification problem. We measure the model's performance with the AUC ROC metric, using the "one-vs-rest" heuristic method, which involves splitting the multiclass dataset into multiple binary classification problems. Furthermore, we calculate the metrics for each class and

compute their average, weighted by the number of true instances for each class.

To extract features from the images, we make use of the ResNet-18 model [13], extracting embeddings from the Average Pooling layer. Since the embedding results in 512 features, which could cause overfitting, we use the mutual information to evaluate the most relevant ones and select the *top K* features, with $K = \sqrt{N}$, where N is the number of data instances in the train set, as suggested in [14].

To evaluate the models' performance across different active learning strategies, we apply a stratified k-fold cross validation [29], using one fold for testing, one fold as a pool of unlabeled data for active learning, and the rest from training the model. We adopt *k=10* based on recommendations by [16], and query all available unlabeled instances to evaluate the active learning approaches. We compare three active learning scenarios: drawing queries through (i) stream-based classifier uncertainty sampling accepting instances with an uncertainty threshold above the 75^{th} percentile of observed instances, (ii) pool-based sampling selecting the instances a given model is most uncertain about, and pool-based sampling considering a query-by-committee strategy, where the committee is created with models trained with the five algorithms we consider in this research: Gaussian Näive Bayes, CART (*Classification and Regression Trees*, similar to C4.5, but it does not compute rule sets), Linear SVM, MLP, and kNN. Comparing deep learning models remains a subject of future work. Finally, we compare the performance of the active learning scenarios computing the average AUC ROC of each fold and assess if the results differences obtained from each model are statistically significant by using the Wilcoxon signed-rank test[26], using a p-value of 0.05.

5 RESULTS AND ANALYSIS

The results obtained from the experiments we ran, and described in Section 4, are presented in Table 1, and Table 2. Table 1 describes the average AUC ROC per each active learning scenario and model for each cross-validation test fold. We observe that the best model across strategies is the MLP, which achieved the best or second-best performance across almost every fold in pool-based and stream-based active learning. Among those two scenarios, the best results were obtained for stream-based active learning. We observed the same across the rest of the models, though the differences were not significant for all but the Näive Bayes models (see Table 2). Query-by-committee displayed a strong performance, showing best results immediately after the MLP. When assessing the statistical significance between the query-by-committee scenario and results obtained from different models with stream-based and pool-based strategies, we observed that differences were significant in all cases, except for the SVM models. SVM models, most widely used in active learning literature related to automated defect inspection, were the third-best models among the tested ones, immediately after the MLPs in stream-based and pool-based active learning and the query-by-committee approach. SVM models did not display significant differences when compared across different active learning scenarios. The worst results were consistently observed for the CART models.

When analyzing the results, we were interested in how the models' performance evolved through time and significant variations between the first and last results observed. To that end, we assessed the statistical significance between the means of the first and last quartiles of the test fold for each active learning scenario. We assessed the statistical significance using the Wilcoxon signed-rank test, with a p-value of 0.05. While such variations existed and were positive in most test folds (the models learned through time), the improvements were not statistically significant in none of the scenarios.

6 CONCLUSION

In this paper, we compared three active learning scenarios (pool-based, stream-based with classifier uncertainty sampling, and query-by-committee) across five machine learning algorithms (Gaussian Näive Bayes, CART, Linear SVM, MLP, and kNN). We found that the best performance was achieved by the MLP model regardless of the active learning strategy. The second-best performance was obtained through the query-by-committee strategy, while the frequently used SVM models ranked third. We found no significant difference between using pool-based or stream-based active learning approaches. Results from the query-by-committee approach were statistically significant in all cases and better than all the models, except for the MLPs. Finally, we found no case where the improvement between the first and last quartile of the test fold in each active learning scenario would be significant. We believe that further investigation is required to determine if a larger pool of unlabeled images would help us achieve such a significant difference. Future work will focus on data augmentation techniques that could help achieve a statistically significant improvement over time when applying active learning techniques.

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 Jelle Keizer and Yvo van Vegten from *Philips Consumer Lifestyle BV*.

REFERENCES

[1] Carlos Beltrán-González, Matteo Bustreo, and Alessio Del Bue. 2020. External and internal quality inspection of aerospace components. In *2020 IEEE 7th International Workshop on Metrology for AeroSpace (MetroAeroSpace)*. IEEE, 351–355.

[2] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9368–9377.

[3] Tajeddine Benbarrad, Marouane Salhaoui, Soukaina Bakhat Kenitar, and Mounir Arioua. 2021. Intelligent machine vision model for defective product inspection based on machine learning. *Journal of Sensor and Actuator Networks* 10, 1 (2021), 7.

[4] Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (1997), 1145 – 1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[5] Amal Chouchene, Adriana Carvalho, Tânia M Lima, Fernando Charrua-Santos, Gerardo J Osório, and Walid Barhoumi. 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)*. IEEE, 127–131.

[6] David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning* 15, 2 (1994), 201–221.

[7] Antoine Cordier, Deepan Das, and Pierre Gutierrez. 2021. Active learning using weakly supervised signals for quality inspection. *arXiv preprint arXiv:2104.02973*

Active Learning scenario	Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
stream-based	CART	0,8168	0,7828	0,7810	0,7694	0,8196	0,7805	0,7843	0,7970	0,8409	0,7940
	kNN	0,9289	0,9121	0,9174	0,8686	0,9024	0,9000	0,9051	0,8960	0,9282	0,9082
	MLP	**0,9900**	**0,9928**	**0,9846**	**0,9563**	**0,9804**	**0,9807**	**0,9710**	**0,9729**	0,9793	**0,9845**
	Näive Bayes	0,8818	0,8668	0,8819	0,8686	0,8829	0,8899	0,8650	0,8877	0,8864	0,9098
	SVM	0,9752	0,9828	0,9725	*0,9530*	0,9816	0,9720	0,9570	0,9412	0,9824	0,9712
pool-based	CART	0,7584	0,7904	0,7543	0,7468	0,8441	0,7730	0,8044	0,7701	0,7850	0,7412
	kNN	0,9189	0,9149	0,9161	0,8581	0,9055	0,9036	0,8961	0,8910	0,9224	0,9056
	MLP	*0,9892*	*0,9921*	*0,9845*	**0,9563**	0,9790	**0,9803**	**0,9702**	**0,9723**	0,9806	*0,9840*
	Näive Bayes	0,8800	0,8654	0,8809	0,8677	0,8813	0,8895	0,8637	0,8873	0,8850	0,9090
	SVM	0,9752	0,9819	0,9726	0,9518	*0,9806*	0,9712	0,9562	0,9412	*0,9823*	0,9722
query-by-committee		0,9774	0,9824	0,9714	0,9500	0,9723	*0,9726*	*0,9597*	*0,9571*	**0,9830**	0,9734

Table 1: AUC ROC values were obtained across the ten cross-validation folds. Best results are bolded, second-best results are highlighted in italics.

Model	Active Learning scenarios		
	stream-based vs. pool-based	stream-based vs. query-by-committee	pool-based vs. query-by-committee
CART	0,0840	**0,0020**	**0,0020**
kNN	0,1309	**0,0020**	**0,0020**
MLP	0,0856	**0,0039**	**0,0039**
Näive Bayes	**0,0020**	**0,0020**	**0,0020**
SVM	0,1824	0,4316	0,6250

Table 2: p-values obtained for the Wilcoxon signed-rank test when comparing the average of AUC ROC results across ten cross-validation folds.

(2021).
[8] Wenting Dai, Abdul Mujeeb, Marius Erdt, and Alexei Sourin. 2018. Towards automatic optical inspection of soldering defects. In *2018 International Conference on Cyberworlds (CW)*. IEEE, 375–382.
[9] Guifang Duan, Hongcui Wang, Zhenyu Liu, and Yen-Wei Chen. 2012. A machine learning-based framework for automatic visual inspection of microdrill bits in PCB production. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 1679–1689.
[10] Tobias Glasmachers. 2017. Limits of end-to-end learning. In *Asian Conference on Machine Learning*. PMLR, 17–32.
[11] Christian Gobert, Edward W Reutzel, Jan Petrich, Abdalla R Nassar, and Shashi Phoha. 2018. Application of supervised machine learning for defect detection during metallic powder bed fusion additive manufacturing using high resolution imaging. *Additive Manufacturing* 21 (2018), 517–528.
[12] Irlán Grangel-González. 2019. *A knowledge graph based integration approach for industry 4.0.* Ph.D. Dissertation. Universitäts-und Landesbibliothek Bonn.
[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[14] Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh, and Edward R Dougherty. 2005. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* 21, 8 (2005), 1509–1515.
[15] Carla Iglesias, Javier Martínez, and Javier Taboada. 2018. Automated vision system for quality inspection of slate slabs. *Computers in Industry* 99 (2018), 119–129.
[16] Max Kuhn, Kjell Johnson, et al. 2013. *Applied predictive modeling*. Vol. 26. Springer.
[17] David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*. Elsevier, 148–156.
[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.
[19] Lingbin Meng, Brandon McWilliams, William Jarosinski, Hye-Yeong Park, Yeon-Gil Jung, Jehyun Lee, and Jing Zhang. 2020. Machine learning in additive manufacturing: A review. *Jom* 72, 6 (2020), 2363–2377.
[20] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. 2018. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–36.
[21] S Ravikumar, KI Ramachandran, and V Sugumaran. 2011. Machine learning approach for automated visual inspection of machine components. *Expert systems with applications* 38, 4 (2011), 3260–3266.
[22] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2020. A survey of deep active learning. *arXiv preprint arXiv:2009.00236* (2020).
[23] Judi E See. 2012. Visual inspection: a review of the literature. *Sandia Report SAND2012-8590, Sandia National Laboratories, Albuquerque, New Mexico* (2012).
[24] Burr Settles. 2009. Active learning literature survey. (2009).
[25] Karin van Garderen. 2018. Active Learning for Overlay Prediction in Semiconductor Manufacturing. (2018).
[26] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*. Springer, 196–202.
[27] Thorsten Wuest, Christopher Irgens, and Klaus-Dieter Thoben. 2014. An approach to monitoring quality in manufacturing using supervised machine learning on product state data. *Journal of Intelligent Manufacturing* 25, 5 (2014), 1167–1180.
[28] Jing Yang, Shaobo Li, Zheng Wang, Hao Dong, Jun Wang, and Shihao Tang. 2020. Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges. *Materials* 13, 24 (2020), 5755.
[29] Xinchuan Zeng and Tony R Martinez. 2000. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* 12, 1 (2000), 1–12.