




# Active learning and novel model calibration measurements for automated visual inspection in manufacturing

Jože M. Rožanec<sup>1,2,5</sup>  · Luka Bizjak<sup>2</sup> · Elena Trajkova<sup>3</sup> · Patrik Zajec<sup>2</sup> · Jelle Keizer<sup>4</sup> · Blaž Fortuna<sup>5</sup> · Dunja Mladenić<sup>2</sup>

Received: 14 September 2022 / Accepted: 16 February 2023 / Published online: 16 March 2023  
© The Author(s) 2023

## Abstract

Quality control is a crucial activity performed by manufacturing enterprises to ensure that their products meet quality standards and avoid potential damage to the brand's reputation. The decreased cost of sensors and connectivity enabled increasing digitalization of manufacturing. In addition, artificial intelligence enables higher degrees of automation, reducing overall costs and time required for defect inspection. This research compares three active learning approaches, having single and multiple oracles, to visual inspection. Six new metrics are proposed to assess the quality of calibration without the need for ground truth. Furthermore, this research explores whether existing calibrators can improve performance by leveraging an approximate ground truth to enlarge the calibration set. The experiments were performed on real-world data provided by *Philips Consumer Lifestyle BV*. Our results show that the explored active learning settings can reduce the data labeling effort by between three and four percent without detriment to the overall quality goals, considering a threshold of  $p = 0.95$ . Furthermore, the results show that the proposed calibration metrics successfully capture relevant information otherwise available to metrics used up to date only through ground truth data. Therefore, the proposed metrics can be used to estimate the quality of models' probability calibration without committing to a labeling effort to obtain ground truth data.

**Keywords** Active learning · Probability calibration · Artificial intelligence · Machine learning · Smart manufacturing · Automated visual inspection

## Introduction

Quality control is one of the key parts of the manufacturing process, which comprehends inspection, testing, and identification to ensure the manufactured products comply with specific standards and specifications (Kurniati et al., 2015; Wuest et al., 2014; Yang et al., 2020). For example, the inspection tasks aim to determine whether a specific part features assembly integrity, surface finish, and adequate geometric dimensions (Newman & Jain, 1995). In addition, product quality is key to the business since it (i) builds trust with the customers, (ii) boosts customer loyalty, and (iii) reinforces the brand reputation.

---

✉ Jože M. Rožanec  
joze.rozanec@ijs.si

Luka Bizjak  
luka.bizjak@ijs.si

Elena Trajkova  
trajkova.elena.00@gmail.com

Patrik Zajec  
patrik.zajec@ijs.si

Jelle Keizer  
jelle.keizer@philips.com

Blaž Fortuna  
blaz.fortuna@qlector.com

Dunja Mladenić  
dunja.mladenic@ijs.si

<sup>1</sup> Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia

<sup>2</sup> Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

<sup>3</sup> Faculty of Electrical Engineering, University of Ljubljana, Tržaška c. 25, 1000 Ljubljana, Slovenia

<sup>4</sup> Philips Consumer Lifestyle BV, Oliemolenstraat 5, Drachten, The Netherlands

<sup>5</sup> Qlector d.o.o., Rovšnikova 7, 1000 Ljubljana, Slovenia

One such quality inspection activity is the visual inspection, considered a bottleneck activity in some instances (Zheng et al., 2020). Visual inspection is associated with many challenges. Some visual inspections require a substantial amount of reasoning capability, visual abilities, and specialization (Newman & Jain, 1995). Furthermore, reliance on humans to perform such tasks can affect the scalability and quality of the inspection. When considering scalability, human inspection requires training inspectors to develop inspection skills; their inspection execution tends to be slower when compared to machines, they fatigue over time and can become absent at work (due to sickness or other motives) (Selvi & Nasira, 2017; Vergara-Villegas et al., 2014). The quality of inspection is usually affected by the inherent subjectiveness of each human inspector, the task complexity, the job design, the working environment, the inspectors' experience, well-being, and motivation, and the management's support and communication (Cullinane et al., 2013; Kujawińska et al., 2016; See, 2012). Manual visual inspection's scalability and quality shortcomings can be addressed through an automated visual inspection.

Automated visual inspection can be realized with Machine Learning models. Technological advances [e.g., Internet of Things or Artificial Intelligence (Rai et al., 2021; Zheng et al., 2021)], and trends in manufacturing [e.g., the Industry 4.0 and Industry 5.0 paradigms (Barari et al., 2021; Rozanec et al., 2022)] have enabled the timely collection of data and foster the use of machine learning models to automate manufacturing tasks while reshaping the role of the worker (Carvajal Soto et al., 2019; Chouchene et al., 2020). Automated visual inspection was applied in several use cases in the past (Beltrán-González et al., 2020; Duan et al., 2012; Jiang & Wong, 2018; Villalba-Diez et al., 2019). Nevertheless, it is considered that the field is still in its early stages and that artificial intelligence has the potential to revolutionize product inspection (Aggour et al., 2019).

While machine learning models can be trained to determine whether a manufactured piece is defective and do so in an unsupervised or supervised manner, no model is perfect. At least three challenges must be faced: (a) how to improve the models' discriminative capabilities over time, (b) how to calibrate the models' prediction scores into probabilities to enable the use of standardized decision rules (Silva Filho et al., 2021), and (c) how to alleviate the manual labeling effort.

This paper presents our approach to addressing these three challenges as follows. Active learning enhances the classification model to address the first challenge. Pool-based and stream-based settings are compared, considering different active learning sample query strategies across five machine learning algorithms. Platt scaling, a popular probability calibration technique, addresses the second challenge. Finally, two scenarios were considered when addressing the reduction of manual labeling effort: (i) manual inspection of cases

where the machine learning model does not predict with enough confidence and (ii) data labeling to acquire ground truth data for the model calibration. The first scenario was addressed by exploring the usage of multiple oracles and soft labeling to reduce the manual inspection effort. Finally, the second and third scenarios were addressed by approximating the ground truth with models' predictions to calibrate the model. Furthermore, several novel metrics to measure the quality of calibration were proposed. The results confirm that they can measure the quality of such calibration without needing a ground truth.

This work extends our previous research described in paper *Streaming Machine Learning and Online Active Learning for Automated Visual Inspection* (Rožanec et al., 2022). In that paper, research was performed to measure the impact of active learning on streaming algorithms. This paper explored batch and online settings, active learning policies, and oracles. This research overcomes some of the shortcomings of the previous research. First, it does not only consider the models' uncertainty to derive data instances to oracles but also a certain quality acceptance level. Second, it calibrates the machine learning models so that through probability calibration, they issue probabilities rather than predictive scores. Third, it increases the amount of data devoted to active learning to ensure more meaningful results. Finally, it focuses on batch machine learning models (which achieve a greater discriminative performance) and studies them in batch and streaming active learning settings. In addition to the above-mentioned items, multiple metrics were developed to assess the calibration quality of a calibrator. The metrics overcome some shortcomings of widely adopted metrics and enable measuring calibration quality when no ground truth is available. The research was performed on a real-world use case with images provided by *Philips Consumer Lifestyle BV* corporation. The dataset comprises images regarding the printed logo on manufactured shavers. The images are classified into three classes: good prints, double prints, and interrupted prints.

The Area Under the Receiver Operating Characteristic Curve [AUC ROC, see (Bradley, 1997)] was used to evaluate the discriminative capability of the classification models. AUC ROC estimates the quality of the model for all possible cutting thresholds. It is invariant to a priori class probabilities and, therefore, suitable for classification tasks with strong class imbalance. Furthermore, given that the models were evaluated in a multiclass setting, the AUC ROC was computed with a one-vs-rest strategy. Furthermore, the performance of multiple probability calibration approaches was measured through the Estimated Calibration Error (ECE) and several novel metrics proposed in this research.

This paper is organized as follows. Section “[Related work](#)” describes the current state of the art and related works. Section “[Approximate model's probabilities calibra-](#)

tion describes novel metrics proposed for probability calibration and how calibration methods can leverage approximate ground truth to enlarge the calibration set. The main novelty regarding the proposed probabilities calibration metrics is the ability to measure calibration quality without needing a ground truth. Section “Use case” describes the use case, while section “Methodology” provides a detailed description of the methodology followed. Section “Experiments” describes the experiments performed, while section “Results and evaluation” presents and discusses the results obtained. Finally, section “Conclusion” presents the conclusions and outlines future work.

## Related work

This section provides a short overview of three topics relevant to this research: (i) the use of machine learning for quality inspection, (ii) active learning, and (iii) probabilities calibration. The following subsections are devoted to each of them.

### Machine learning for quality inspection

A comprehensive and reliable quality inspection is indispensable to the manufacturing process, and high inspection volumes turn inspection processes into bottlenecks (Schmitt et al., 2020). Machine Learning has been recognized as a technology that can drive the automation of quality inspection tasks in the industry. Multiple authors report applying it for early prediction of manufacturing outcomes, which can help drop a product that will not meet quality expectations and avoid investment in expensive manufacturing stages. Furthermore, similar predictions can be used to determine whether the product can be repaired and therefore avoid either throwing away a piece to which the manufacturing process was invested or selling a defective piece with the corresponding costs for the company (Weiss et al., 2016). Automated visual inspection refers to image processing techniques for quality control, usually applied in the production line of manufacturing industries (Beltrán-González et al., 2020). It has been successfully applied to determine the end quality of the products. It provides many advantages, such as performing non-contact inspection that is not affected by the type of target, surface, or ambient conditions (e.g., temperature) (Park et al., 2016). In addition, visual inspection systems can perform multiple tasks simultaneously, such as object, texture, or shape classification and defect segmentation, among other inspections. Nevertheless, automated visual inspection is a challenging task given that collecting the dataset is usually expensive, and the methods developed for that purpose are dataset-dependent (Ren et al., 2017).

Jian et al. (2017) considers three approaches that exist toward automated visual inspection: (a) classification, (b) background reconstruction and removal (reconstruct and remove background to find defects in the residual image), and (c) template reference (comparing a template image with a test image). Tsai and Lai (2008) describe how TFT-LCD panels and LCD color filters were inspected by comparing surface segments containing complex periodic patterns. Lin et al. (2019) describes how defect inspection on LED chips was automated using deep Convolutional Neural Networks (CNN). Kang and Liu (2005) successfully applied feed-forward networks to detect surface defects on cold-rolled strips. In the same line, Yun et al. (2014) proposed a novel defect detection algorithm for steel wire rods produced by the hot rolling process. Valavanis and Kosmopoulos (2010) compared multiple machine learning models (Support Vector Machine, Neural Network, and K-nearest neighbors (kNN)) on defect detection in weld images. Park et al. (2016) developed a CNN and compared it to multiple models (particle swarm optimization-imperialist competitive algorithm, Gabor-filter, and random forest with variance-of-variance features) to find defects on silicon wafers, solid paint, pearl paint, fabric, stone, and wood surfaces. Furthermore, Aminzadeh and Kurfess (2019) described how Bayesian classification enabled online quality inspection in a powder-bed additive manufacturing setting. Multiple authors developed machine learning algorithms for visual inspection leveraging feature extraction from pre-trained models (Cohen & Hoshen, 2020; Li et al., 2021; Jezek et al., 2021). While much research was devoted to supervised machine learning methods, unsupervised defect detection was explored by many authors, who explored using Fourier transforms to remove regularities and highlight irregularities (defects) (Aiger & Talbot, 2012) or employed autoencoders to find how a reference image differs from the expected pattern (Mujeeb et al., 2018; Zavrtnik et al., 2021, 2022).

### Active learning

Active learning is a subfield of machine learning that studies how an active learner can best identify informative unlabeled instances and requests their labels from 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 are (i) uncertainty sampling (select an unlabeled sample with the highest uncertainty, given a certain metric, or machine-learning model (Lewis & Catlett, 1994)), or (ii) query-by-committee [retrieve the unlabeled sample with the highest disagreement between a set of

forecasting models (*committee*) (Cohn et al., 1994; Settles, 2009)] can be found. More recently, new scenarios have been proposed leveraging reinforcement learning, where an agent learns to select images based on their similarity, and rewards obtained are based on the oracle's feedback (Ren et al., 2020). In addition, it has been demonstrated that ensemble-based active learning can effectively counteract class imbalance through newly labeled image acquisition (Beluch et al., 2018). While active learning reduces the required volume of labeled images, 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 (Cordier et al., 2021).

Active learning was successfully applied in manufacturing, but scientific literature remains scarce on this domain (Meng et al., 2020). Some use cases include the automatic optical inspection of printed circuit boards (Dai et al., 2018), media news recommendation in a demand forecasting setting (Zajec et al., 2021), and the identification of the local displacement between two layers on a chip in the semiconductor industry (van Garderen, 2018).

## Probabilities calibration

Probabilities denote the likelihood that a particular event will occur and are expressed as a real number between zero and one (Cheeseman, 1985). Many machine learning models output prediction scores which cannot be directly interpreted as probabilities. Therefore, such models can be calibrated (mapped to a known scale with known properties), ensuring the prediction scores are converted to probabilities. Probability calibration aims to provide reliable estimates of the true probability that a sample is a member of a class of interest. Such calibration (a) usually does not decrease the classification accuracy, (b) enables using provides thresholds on the decision rules and therefore minimizes the classification error, (c) ensures decision rules and their maximum posterior probability are fully justified from the theoretical point of view, (d) can be easily adapted to changes in class and cost distributions, and therefore (e) is key to decision-making tasks (Cohen & Goldszmidt, 2004; Song et al., 2021).

The  $k$ -class probabilistic classifier is considered well-calibrated if the predicted  $k$ -dimensional probability vector has a distribution that approximates the distribution of the test instances. While a single accepted notion of probabilistic calibration exists for binary classifiers, the definition for multiclass settings has multiple nuances. Three kinds of probability calibration are described in the literature for multiclass settings: (i) *confidence calibration* [aims only to calibrate the classifier's most likely predicted class (Song et al., 2021)], (ii) *class-wise calibration* (attempts to calibrate the scores for each class as marginal probabilities), and (iii) *multi-class calibration* (seeks to create an entire vector of predicted prob-

abilities so that for any prediction vector the proportion of classes among all possible instances getting the same prediction, are equal to the probabilities for those classes in the predicted vector).

Multiple probability calibration methods have been proposed in the scientific literature. The post-hoc techniques aim to learn a calibration map for a machine-learning model based on hold-out validation data. In addition, popular calibration methods for binary classifiers include logistic calibration (Platt scaling), isotonic calibration, Beta calibration, temperature calibration, and binning calibration.

Empirical binning builds the calibration map by computing the empirical frequencies within a set of score intervals. It can therefore capture arbitrary prediction score distributions (Kumar et al., 2019). Isotonic regression computes a regression assuming the uncalibrated model has a set of non-decreasing constant segments corresponding to bins of varying widths. Given its non-parametric nature, it avoids a model misfit, and due to the monotonicity assumption, it can find optimal bin edges. Nevertheless, training times and memory consumption can be high on large datasets and give sub-optimal results if the monotonicity assumption is violated. Platt scaling (Platt, 2000) aims to transform prediction scores into probabilities through a logistic regression model, considering a uniform probability vector as the target. While the implementation is straightforward and the training process is fast, it assumes the input values correspond to a real scalar space and restricts the calibration map to a sigmoid shape. Probability calibration trees evolve the concept of Platt scaling, identifying regions of the input space that lead to poor probability calibration and learning different probability calibration models for those regions, achieving better overall performance (Leathart et al., 2017). Beta calibration was designed for probabilistic classifiers. It assumes that the scores of each class can be approximated with two Beta distributions and is implemented as a bivariate logistic regression. Temperature scaling uses a scalar parameter  $T > 0$  (where  $T$  is considered the temperature) to rescale logit scores before applying a softmax function to achieve recalibrated probabilities with better spread scores between zero and one. It is frequently applied to deep learning models, where the prediction scores are frequently strongly skewed towards one or zero. Furthermore, the method can be applied to generic probabilistic models by transforming the prediction scores with a logit transform (Guo et al., 2017). This enables calculating the score against a reference class and obtaining the ratio against other classes. Nevertheless, the method is not robust in capturing epistemic uncertainty (Ovadia et al., 2019). Finally, the concept of temperature scaling is extended in vector scaling, which considers that a different temperature for each class can be specified, and matrix scaling, which considers a matrix and intercept parameters (Song et al., 2021).

Several metrics and methods were proposed to assess the quality of the calibration. Reliability diagrams plot the observed relative frequency of predicted scores against their values. They, therefore, enable to quickly assess whether the event happens with a relative frequency consistent with the forecasted value (Bröcker & Smith, 2007). On the other hand, validity plots aim to convey the bin frequencies for every bin and therefore provide valuable information regarding miscalibration bounds (Gupta & Ramdas, 2021). Among the metrics, the binary ECE measures the average gap across all bins in a reliability diagram, weighted by the number of instances in each bin, considering the labeled samples of a test set. In the same line, the binary Maximum Calibration Error computes the maximum gap across all bins in a reliability diagram. The Confidence Estimated Calibration Error measures the average difference between accuracy and average confidence across all bins in a confidence reliability diagram, weighted by the number of instances per bin. A different approach is followed by the Brier score, which measures the mean squared difference between the predicted probability and the actual outcome. While the ECE metric is widely accepted, research has shown that it is subject to shortcomings (Nixon et al., 2019; Posocco & Bonnefoy, 2021). One of such shortcomings is that when using fixed calibration ranges, some bins contain most of the data, resulting in the metric's decreased sharpness. Furthermore, ECE is measured across non-empty bins, failing to account for the overall distribution of positives across the mean predicted probabilities. Measuring probabilistic calibration remains a challenge (Nixon et al., 2019).

While many probability calibration methods and metrics have been developed, most of them were conceived considering probability calibration must be done based on some ground truth. Nevertheless, acquiring data for such ground truth is expensive (requires labeled instances), limits the amount of data seen to build such a probability calibration map, and therefore introduces inaccuracies due to the inherent characteristics of the sample. To address this void, this research proposes labeling each predicted data instance according to the predicted class with the highest score or most likely class if the highest predicted scores are equal. Assuming the classifier could perform with perfect discriminative power in the best case, such labels would equal the ground truth. Furthermore, this research proposes metrics to assess the discrepancy between an ideal probability calibration scenario and the calibrated classifier to measure the quality of probability calibration achieved. By doing so, the calibrators' quality over time can be measured without needing any data labeling for such an assessment. Furthermore, it enables exploring approximate model's probabilities calibration, training a calibrator from a ground truth approximated with predicted labels. This idea is further explained

and developed in section “[Approximate model's probabilities calibration](#)”.

## Approximate model's probabilities calibration

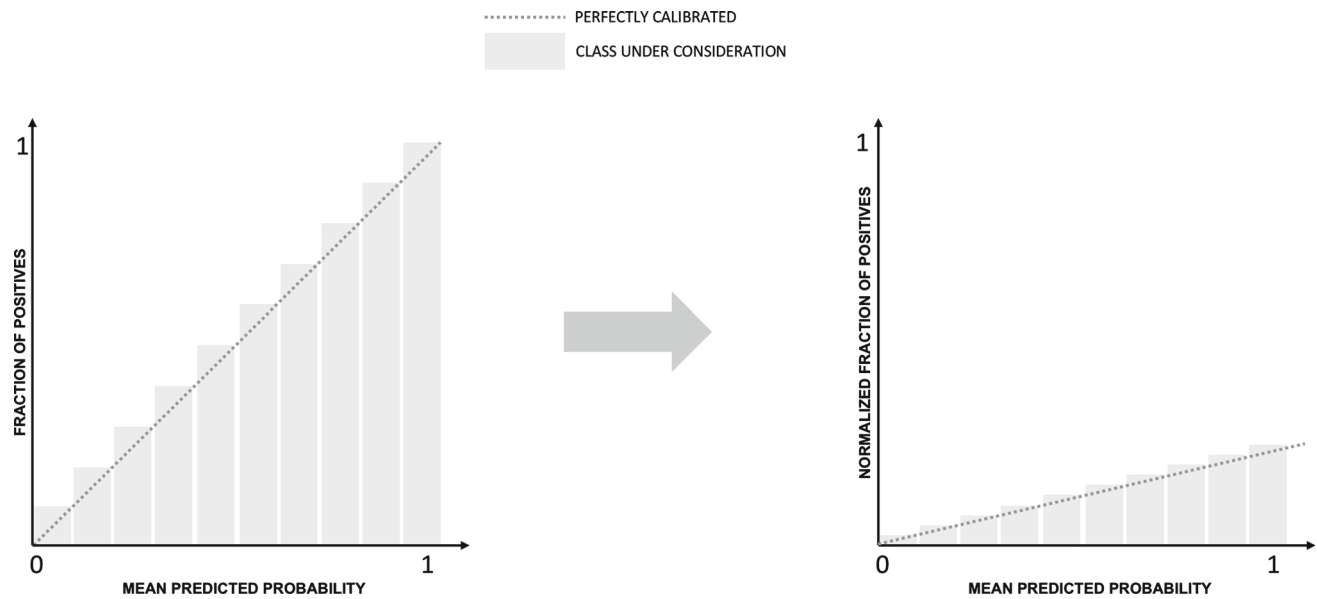
### Towards approximate probability calibration models

This research proposes metrics and an approach to calibrating machine learning prediction scores to probabilities using a ground truth approximation. The approach considers building an initial calibration set, as it is common practice for probability calibration methods. A calibration set has (a) several prediction scores used to perform the probability calibration and (b) the ground truth labels for the corresponding data instances. Using both, a mapping is created between the prediction scores and the probability of a class outcome. Nevertheless, the limited amount of data in the calibration set can impact the fidelity of the calibration. In particular, the distribution of predictive scores between the calibration set and the predictions performed in a production environment can differ.

The final prediction of a calibrated model has at least two sources of error: (a) the classification model, which does not perfectly predict the target class, and (b) the probability calibration technique, which does not produce a perfect probabilistic mapping between the predicted scores and the target class. While (a) directly affects the refinement loss (loss produced by assigning the same probability for instances from different classes), (b) affects the calibration loss (loss due to the difference between the predicted probabilities and observed positive instances for such an output). While metrics and plots exist to assess the quality of the probability calibration, such means require a ground truth to evaluate the probability calibration. While the requirement for a ground truth allows for an exact estimate of the classifier on that particular hold-out data, it has at least two drawbacks: (i) it requires labeling a certain amount of data to perform the evaluation, and (ii) such data may not be representative of current or future data distributions observed in a production environment.

### Intuitions behind a calibration without a ground truth

Current scientific literature considers the quality of a model calibration can be measured by comparing, given a fixed class, whether the fraction of positives does correspond to the predicted mean probability of a given classifier. The fraction of positives empirically measures the likelihood of positive class events for the class under consideration within a specific mean probability range (bin). In a well-calibrated model,



**Fig. 1** The figure presents two calibration plots. On the left, the calibration plot shows a perfectly calibrated calibrator (where the fraction of positives for the class under consideration equals the mean predicted

probability). On the right, the same information is presented, but normalizing the values of the plot on the right to ensure the sum of their values equals one

the likelihood of the occurrence of positive class events in a particular bin for the class under consideration matches the mean predicted probability, revealing a linear relationship between the mean predicted probability and the likelihood of the occurrence in that bin of the positive class event for the class considered (see Fig. 1). Furthermore, a perfectly calibrated classifier is only possible for a binary classification problem with no class imbalance. Class imbalance or multiple classes introduce distortions regarding the frequency with which the positive class is observed within a given predicted mean probability range compared to the frequency with the other events occurring within that mean probability range.

For a well-calibrated classifier, each of the predicted classes is expected to behave as shown in Fig. 1. Therefore, while class imbalance or a multi-class setting can introduce distortions to the histogram's shape, the distance to the ideal case could be measured by comparing the histogram shape of a perfectly calibrated model for a given class and the shape of the histogram in the real-world case under consideration. To estimate how close the histograms are from each other, optimal transport is used (Peyré et al., 2019; Villani, 2009). In particular, the Wasserstein distance measures the distance between the two histogram distributions. We consider the Wasserstein distance between the histograms representing the existing calibration and a perfect one. The distance denotes the improvement opportunity regarding the specific calibration model to achieve a perfect calibration (or a desired calibration according to the reference histogram). Nevertheless, the fraction of positives for a given class cannot

be computed when no ground truth is available. Therefore, we reframe the problem so that the goodness of a model calibration can be evaluated even without considering a ground truth.

Considering the information available in Fig. 1 and a particular class  $j$ , and considering each prediction regarding class  $j$  an event  $x$ , we are interested on two types of events:  $E_1 = \{x \text{ corresponds to bin } i\}$ , and  $E_2 = \{x \text{ corresponds to class } j\}$ . Furthermore, we are interested in calibrating the model so that the resulting score indicates  $p_j(E_2|E_1)$ .

### Intuition 1: Considering a perfectly calibrated classifier

Let us consider the case of a perfectly calibrated classifier. Given a perfectly calibrated classifier, the fraction of positives for a given class must match the mean predicted probability. The fraction of positives within a certain bin  $i$  can be considered the empirical computation of  $p_j(E_2|E_1)$ .  $E_1$  and  $E_2$  are not independent events, given the probability of belonging to class  $j$  should be higher in bins representing a higher mean predicted probability. Therefore,  $p_j(E_2|E_1) = \frac{p_j(E_2 \cap E_1)}{p_j(E_1)}$ . Considering a balanced binary classification problem, the number of predictions issued for each mean predicted probability range must be equal to verify the symmetry regarding the fraction of positives observed in the mean probability ranges for both classes. Fluctuations regarding the fraction of positives observed in the mean predicted probability ranges translate into an unequal number of

predictions in them and directly impact the quality of the calibration. Based on this observation, given the abovementioned equation,  $p_j(E_2|E_1) = \frac{p_j(E_2 \cap E_1)}{p_j(E_1)}$ ,  $p_j(E_1)$  is constant, and can be empirically computed as  $p_j(E_1) = \frac{1}{\# \text{ of bins}}$ . The number of predictions for a given class  $j$  is computed as the count of predictions where the highest predicted value was issued for that class  $j$ . While  $p_j(E_2 \cap E_1)$  cannot be computed without ground truth, the expected values that must be satisfied for each bin for  $p_j(E_2|E_1)$  are known. Therefore, we envision at least two ways to estimate the mismatch between the ideal case and the case under consideration. First, the value of  $p_j(E_2 \cap E_1)$  can be inferred based on the expected  $p_j(E_2|E_1)$  for a particular bin and the empirical computation of  $p_j(E_1)$  to then measure the Wasserstein distance between the resulting distributions. Second, it could be estimated by only considering  $p_j(E_1)$  and measuring the Wasserstein distance between the ideal distribution (an equal number of predictions per mean predicted probability range) and the distribution of predictions obtained from the calibrated classifier under consideration (number of predictions per bin, that are empirically measured—usually the amount of predictions is not equal across bins given the calibrated classifier’s imperfection). Each class’s calibration quality could be estimated in both cases by comparing two histograms: the ideal case and the calibration model under consideration. The distance between both distributions computes measures how far the particular calibrator is from a perfectly calibrated case.

While the case above was demonstrated for a balanced binary classification problem, it approximately holds for multiclass settings and cases with class imbalance. In these scenarios, we aim to calibrate each class as perfectly as possible, even though a perfect calibration cannot be achieved. Nevertheless, how well-calibrated each class is against the ideal case can still be assessed by comparing the distributions described above.

### Intuition 2: Considering a perfect classifier

Let us consider the case of a perfect classifier. Given a perfect classifier, the prediction equals the ground truth regarding a positive class event for the class under consideration. Therefore, two scenarios are considered: (a) degrade the classifiers’ performance to achieve a calibrated classifier, or (b) spread the predicted values within a specific range so that they emulate particular calibration. It must be noted that while (a) can still satisfy the definition of probability considered for calibration, (b) does not.

For (a), the classifier’s performance must be degraded due to the inherent definition of probabilities used in this problem: the calibration model will ensure a proportion of positive events regarding a class given a mean predicted probability

bin. Therefore, given  $n = \text{number of classes}$ , the highest predicted value for each class will not issue only data instances of that class above  $1/n$ . Furthermore, some cases will be lost under the  $1/n$  threshold.

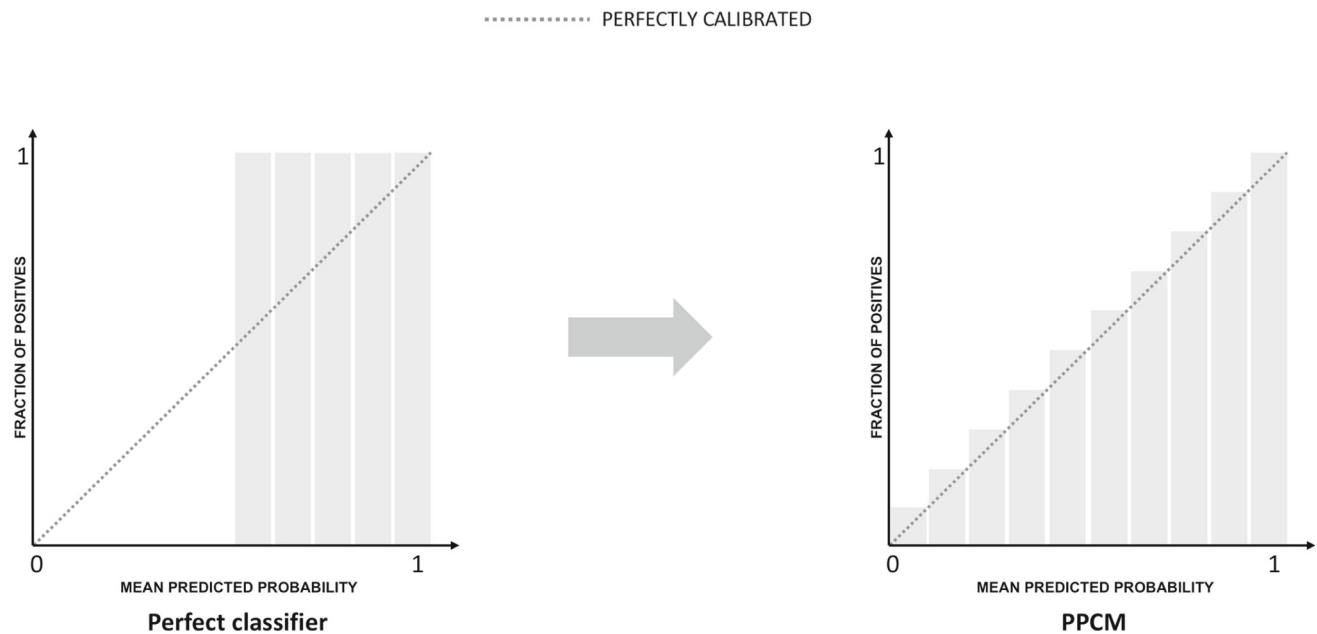
On the other hand, for (b), the abovementioned equation  $p_j(E_2|E_1) = \frac{p_j(E_2 \cap E_1)}{p_j(E_1)}$  can be considered. It is known that for a perfect classifier, the following is true:  $p_j(E_2) = 0$  or  $p_j(E_2) = 1$ . Furthermore,  $E_1$  and  $E_2$  can be considered dependent events, given  $p_j(E_2) = 0$  for bins below a certain threshold, and  $p_j(E_2) = 1$  otherwise (see Fig. 2). In addition, the mean predicted probability would not match the fraction of positives, given the classifier is perfect: each prediction perfectly identifies the target class. Therefore, this scenario *per se* violates the idea behind probabilities calibration. Nevertheless, the best approximation towards Fig. 1 would be to achieve an increasing number of predictions per mean predicted probability range (histogram bin) for a specific class. To avoid degrading the models’ discriminative power, such a mapping function will not issue scores below  $1/n$ , where  $n = \text{number of classes}$ .

## Intuitions materialized

### From intuitions to approximate calibrators

To perform model calibration, a function that can map the predictive scores of a machine-learning model to probability scores is required. Ideally, such probability scores would indicate  $p_j(E_2|E_1)$ . When no ground truth is available, the intuitions described above can be considered to reproduce some scenarios where the resulting probability score distribution can be compared against an ideal probability score distribution. Therefore, we consider labeling the predicted data instances with the class with the highest predicted score. In case two classes hold equal scores, we decide on the most probable one based on the class imbalance observed in the train test. For balanced datasets, the class can be assigned randomly, given no other information exists to guide the decision. The more perfect the classification model, the closer will the assigned labels be to the ground truth. Given data instances with predicted scores and assigned labels, a calibrator can be fitted to map the classifier’s output to a calibrated probability.

To realize the abovementioned calibration without ground truth, at least the following preconditions must be met: (a) no concept drift exists, (b) no covariate shift exists, and (c) the values of the features in the production environment remain within the ranges considered when training the machine learning model.



**Fig. 2** The figure presents two calibration plots. On the left, the calibration plot shows a perfect binary classifier, while on the right we find a perfectly calibrated binary classifier

### From intuitions to metrics

In Sections “[Intuition 1: Considering a perfectly calibrated classifier](#)” and “[Intuition 2: Considering a perfect classifier](#)”, the cases of a perfectly calibrated model and a perfect classifier were considered. While in the case of a perfect classifier, a ground truth is not needed (the predicted labels equal the ground truth), non-perfect classifiers approximate such a ground truth to a certain degree (measured as the classifiers’ performance). Furthermore, regardless of the calibration technique, it was shown that a certain correlation between the calibration quality and the calibration score distribution exists. In particular, it was shown that for each class  $k$  a histogram could be computed showing (a) the number of predictions per bin and (b) the proportion of positive class occurrences per mean predicted probability bin. Both could then be compared against ideal cases. A certain advantage of (a) is that it does not require ground truth or ground truth approximation to determine whether some bins are under or over-assigned. While such an imbalance certainly signals a calibration error, the histogram lack information regarding the composition of each bin. In particular, they provide no information on whether the positive class occurrences increase according to the value of the mean predicted probability bin. This can only be measured in (b), comparing all cases against an ideal calibration histogram. For multiclass problems, each class could be compared against such a histogram, and the resulting scores averaged (Fig. 3).

To estimate how close a probability calibration method is w.r.t. the target (ideal) histogram, optimal transport is used

(Peyré et al., 2019; Villani, 2009). In particular, the Wasserstein distance between two histogram distributions is considered: a histogram constructed with the calibrator scores and a histogram corresponding to the ideal scenario. Based on them, we propose a metric that can be used to estimate the quality of calibration of any calibrator given certain ground truth. We name it *Probability Calibration Score* (PCS—see Eq. 1). The proposed metrics issue a value between zero and one: PCS is zero when the model is not calibrated and one when the model is perfectly calibrated. Furthermore, a weighted metric variant can also be considered (wPCS—see Eq. 2), where the proportion of each class among the observed instances weights the Wasserstein distances.

$W_1(h_i, h_{ref})$  is the 1-Wasserstein distance between the histogram  $h_i$  and the reference histogram  $h_{ref}$  and  $n$  is the number of classes.

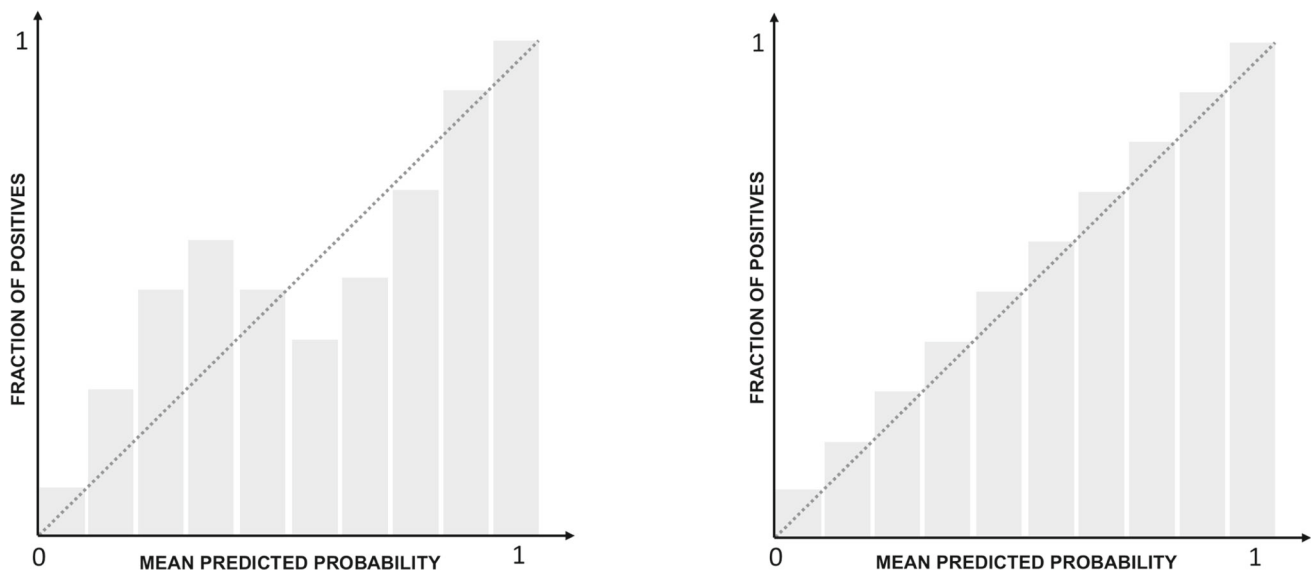
$$PCS = \sum_{i=1}^n \frac{1 - W_1(h_i, h_{ref})}{n} \quad (1)$$

$W_1(h_i, h_{ref})$  is the 1-Wasserstein distance between the histogram  $h_i$  and the reference histogram  $h_{ref}$  and  $w_i$  is the weight of a particular class.  $n$  indicates the number of classes under consideration.

$$wPCS = \sum_{i=1}^n (1 - W_1(h_i, h_{ref})) \cdot w_i \quad (2)$$

To ensure the histograms are comparable, they are normalized, ensuring that the sum of their values equals one. To ensure the Wasserstein distance remains between zero





**Fig. 3** The figure illustrates two sample histograms: the histogram on the left corresponds to some sub-optimally calibrated classifier. In contrast, the histogram on the right (reference histogram) corresponds to a perfectly calibrated classifier

and one, the distance between both distributions is divided by the distance measured between the worst-case scenario and the reference ideal histogram (see Fig. 4). In Fig. 4, we consider the Wasserstein distance between the case on the left and the distribution of a Perfect Probability Calibration Model (PPCM) to be the highest among possible calibration scenarios.

When assessing the performance of an approximate calibrated model, two errors must be taken into account: (i) the classification error, given the classifier does not perfectly predict the target class (and the ground truth is approximated with such predictions), and (ii) the probability calibration technique, which does not produce a perfect probabilistic mapping between the predicted scores and the (approximated) target class. To measure (i), we choose the AUC ROC metric, which is not affected by the class imbalance. AUC ROC can be computed in a multiclass setting with a one-vs-rest or one-vs-one strategy. We measure it on the test set. We consider (ii) can be measured using the Wasserstein distance, comparing the ideal calibration histogram and a histogram where the proportion of positive class occurrences (given the approximate ground truth) is considered per mean predicted probability bin.

We propose two metrics, which we name Additive Probability Calibration Score (APCS—see Eq. 6) and Multiplicative Probability Calibration Score (MPCS—see Eq. 8). Both summarize the calibrated models’ performance, considering the classifier’s imperfection (see Eq. 3) and the calibration error incurred due to the lack of ground truth. To ensure the Wasserstein distance remains between zero and one, we compute a normalized histogram, ensuring the area of the entire histogram equals one. The proposed metrics issue a

value between zero and one, and in both cases, the higher the value, the better the model. Furthermore, we also provide a weighted version of both metrics (wAPCS (see Eq. 7) and wMPCS (see Eq. 9)), which aim to weight the Wasserstein distance between the normalized histograms obtained from a calibrator and the ideal histogram with the class weights (see Eqs. 4 and 5 for  $APCS_W$  and  $wAPCS_W$ , and Eqs. 1 and 2 for MPCS and wMPCS).

APCS is zero when the model has no discriminative power and is not calibrated, and one when the model is perfectly calibrated and shows no classification error on the test set. The APCS metric is detailed in Eq. 6.

$K$  is used to measure classifiers’ discriminative power.  $AUC_{ROC_{Classifier_{test}}}$  corresponds to the classifiers’ AUC ROC measured on the test set.

$$K_{AUC_{ROC}} = |0.5 - AUC_{ROC_{Classifier_{test}}}| \tag{3}$$

Component for Wasserstein distance measurement between an ideal calibrator and the calibrator under consideration, as used for the APCS metric.

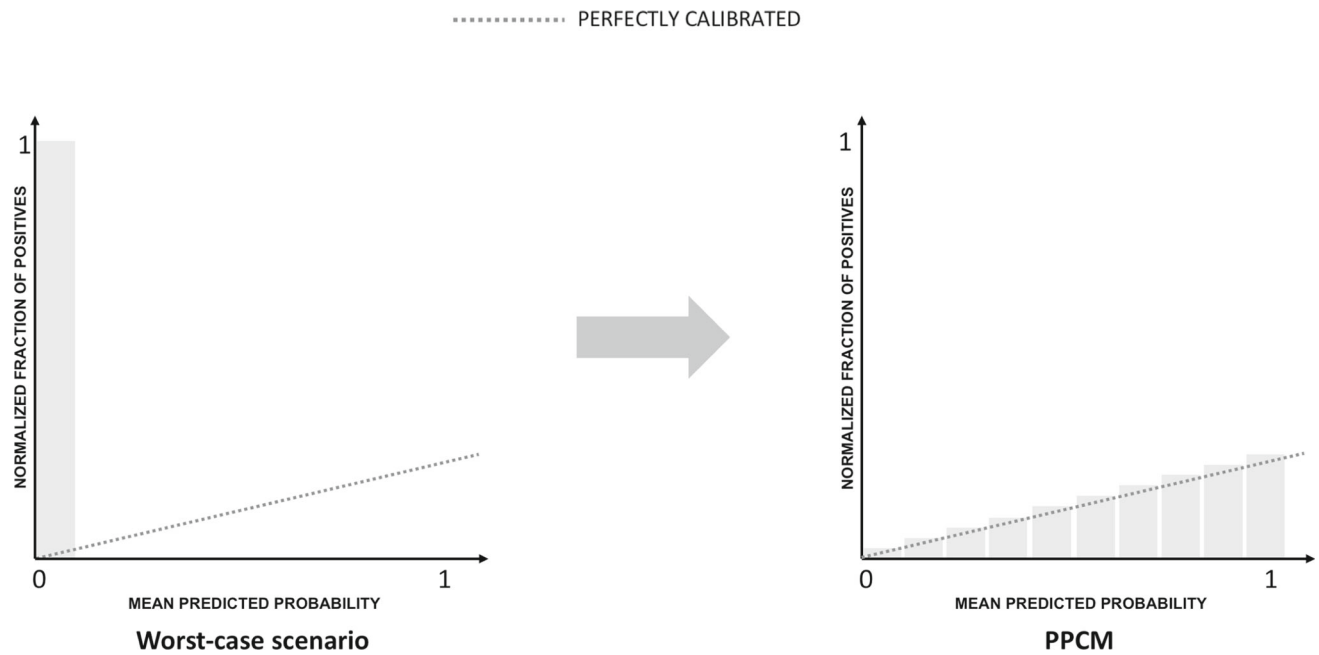
$$APCS_W = 0.5 \cdot PCS \tag{4}$$

Component for Wasserstein distance measurement between an ideal calibrator and the calibrator under consideration, as used for the wAPCS metric.

$$wAPCS_W = 0.5 \cdot wPCS \tag{5}$$

APCS metric definition.

$$APCS = K_{AUC_{ROC}} + APCS_W \tag{6}$$



**Fig. 4** The figure illustrates two sample calibration plots: the calibration plot on the left corresponds to a calibrated classifier where all positives were assigned to a zero mean predicted probability (worst-case scenario). In contrast, the calibration plot on the right (reference

histogram) corresponds to a perfectly calibrated classifier. Both calibration plots correspond to normalized cases, where the sum of the values equals one

wAPCS metric definition.

$$wAPCS = K_{AUCROC} + wAPCS_W \tag{7}$$

On the other hand, MPCS and wMPCS correspond to zero when (a) the classifiers’ predictive ability is no better than random guessing or (b) the Wasserstein distance between histograms is highest (equal to one). Moreover, MPCS and wMPCS correspond to one when (a) the classifiers’ predictive ability is perfect, and (b) the calibration is perfect w.r.t. the target histogram  $h$  of choice. The MPCS metric is detailed in Eq. 8.

MPCS metric definition

$$MPCS = K_{AUCROC} \cdot PCS \tag{8}$$

MPCS metric definition

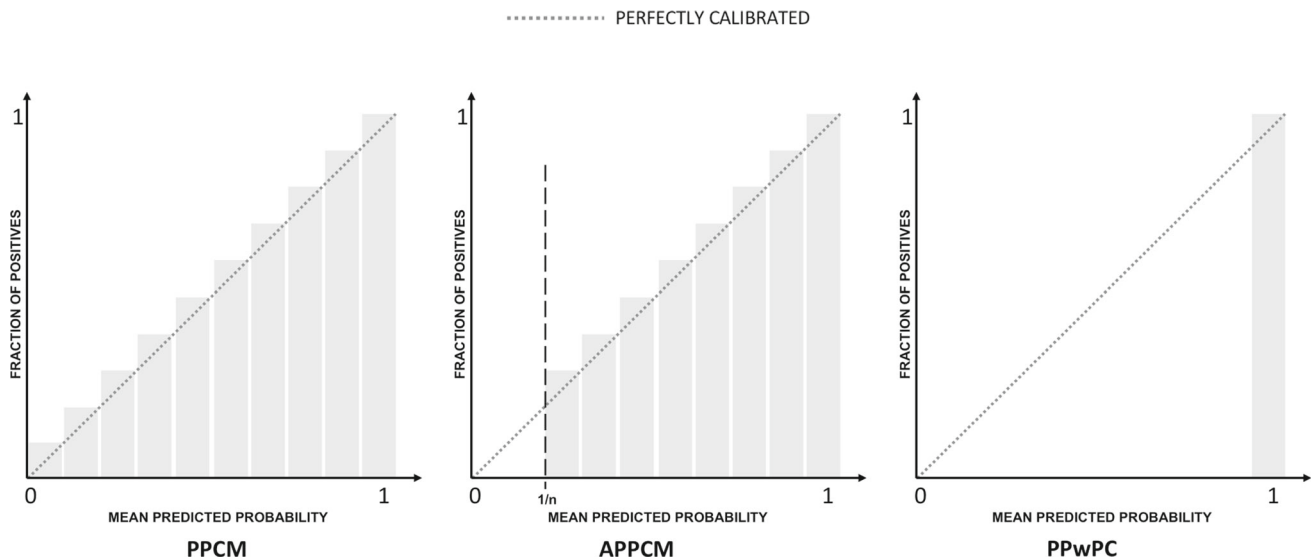
$$wMPCS = K_{AUCROC} \cdot wPCS \tag{9}$$

For models’ probability calibration, PCS, APCS, and MPCS assume an ideal reference histogram. Three histograms are presented in Fig. 5 corresponding to (a) a Perfect Probability Calibration Model (PPCM), (b) an Almost Perfect Probability Calibration Model (APPCM), and (c) Perfect Classification with Perfect Confidence (PPwPC). While only PPCM can be used for strict probability calibration, the other

two reference histograms measure how far the distributions of the predicted values are from other desired distribution shapes. In particular, APPCM achieves a similar spread of predicted probabilities as PPCM but neglects the segment of predictions below  $1/n$  (with  $n = \text{number of classes}$ ), where the classifier would become suboptimal. On the other hand, PPwPC advocates for a classifier where all scores are pushed toward the highest possible score for a given class. This research only considers the PPCM reference histogram to compute the above-described metrics.

### Use case

*Philips Consumer Lifestyle BV* in Drachten, The Netherlands, is one of Philips’ biggest development and production centers in Europe. They use cutting-edge production technology to manufacture products ceaselessly. One of their improvement opportunities is related to visual inspection, where they aim to identify when the company logo is not properly printed on the manufactured products. They have multiple printing pad machines, from which the products are handled and inspected on their visual quality and removed if any error is detected. Experts estimate that a fully automated procedure would speed up the process by more than 40%. Currently, there are two defects associated with the printing quality of the logo (see Fig. 6): double prints (the whole logo



**Fig. 5** The figure illustrates three histograms that correspond to ideal cases described in this section: Perfect Probability Calibration Model (PPCM), Almost Perfect Probability Calibration Model (APPCM), and Perfect Classification with Perfect Confidence (PPwPC)



**Fig. 6** The images shown above correspond to three possible classes: good (no defect), double print (defective), and interrupted print (defective)

is printed twice with a varying overlap degree) and interrupted prints (the logo displays small non-pigmented areas, similar to scratches).

Machine learning models can be developed to automate the visual inspection procedure (Rippel et al., 2021; Zavrzanik et al., 2022). However, given that such models are imperfect, the manual revision can be used as a fallback to inspect the products about which the uncertainty of the machine learning model exceeds a certain threshold. Such decisions can be made based on simple decision rules, quality policies, and the probability of obtaining a defective product given a particular prediction score. Furthermore, products sent for manual inspection can be prioritized using different criteria to enhance the existing defect detection machine learning model. This research explores the abovementioned capabilities through multiple experiments, building supervised models, leveraging active learning, and comparing six machine learning algorithms. Furthermore, new measures for probability calibration are explored, and experiments are executed to determine whether existing calibration techniques would benefit from enlarging the calibration set with approximate ground truth. The experiments were conducted on a

dataset of 3518 labeled images, all corresponding to manufactured shavers.

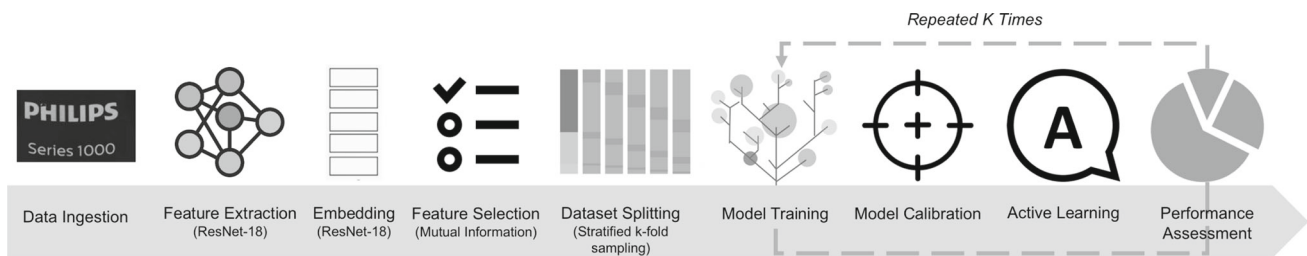
## Methodology

The research presented in this paper was performed using the Python language, and open source libraries, such as scikit-learn (Buitinck et al., 2013) and netcal (Küppers et al., 2020).

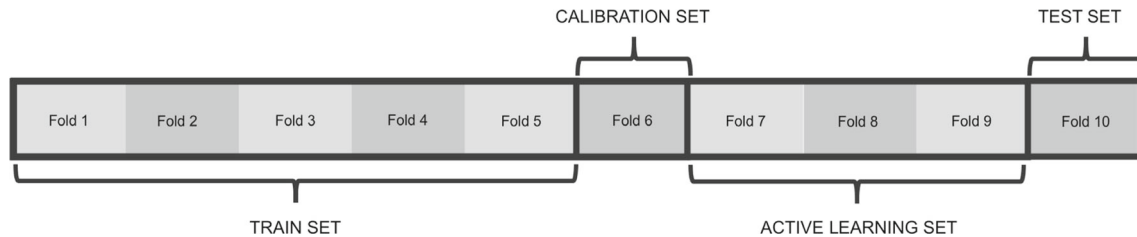
### Methodological aspects to evaluate active learning strategies

We frame the automated defect detection as a supervised, multiclass classification problem. A ResNet-18 model (He et al., 2016) was used for feature extraction. 512 values long vectors were extracted for each image obtained from the average pooling layer. To avoid overfitting, the procedure suggested by Hua et al. (2005) was followed by selecting the *top K* features, with  $K = \sqrt{N}$ , where  $N$  is the number of data instances in the train set. Features' relevance was assessed considering the mutual information score, which measures any relationship between random variables. It is considered that the mutual information score is not sensitive to feature transformations if these transformations are invertible and differentiable in the feature space or preserve the order of the original elements of the feature vectors (Vergara & Estévez, 2014) (Fig. 7).

To evaluate the models' and active learning scenarios' performance, a stratified  $k$ -fold cross validation (Zeng & Martinez, 2000) was applied, considering  $k=10$  based on recommendations by Kuhn and Johnson (2013). One fold



**Fig. 7** The methodology we followed to train and assess machine learning models and active learning scenarios



**Fig. 8** A 10-fold stratified cross-validation was used. The dataset was split for four purposes: train, test, probabilities calibration, and simulate unlabeled data under an active learning setting

was used for testing (*test set*), and one for machine learning models' probabilities calibration (*calibration set*). Three folds were used to simulate a pool of unlabeled data for active learning (*active learning set*), and the rest to train the model (*train set*) (see Fig. 8). Samples are selected from the *active learning set* to be annotated by the oracle and then added to the training set, on which the models are retrained. In this research, two types of oracles were considered: (a) machine oracles, which can be imperfect, and (b) human annotators (assumed to be ideal). Five machine learning algorithms were evaluated: Gaussian Naïve Bayes, CART (*Classification and Regression Trees*, similar to C4.5, but it does not compute rule sets), Linear SVM, kNN, and Multilayer perceptron (MLP).

To evaluate the discriminative power of the machine learning models and how it is enhanced over time through active learning, the AUC ROC metric was computed. Given the multiclass setting, the “one-vs-rest” heuristic was selected, splitting the multiclass dataset into multiple binary classification problems and computing their average, weighted by the number of true instances for each class. In addition, to assess the usefulness of the active learning approaches, the AUC ROC values obtained by evaluating the model against the test fold for the first (Q1) and last (Q4) quartiles of instances queried in an active learning setting were compared. The amount of manual work saved under each active learning setting and the soft-labeling approaches' precision were also evaluated.

Through different experiments (detailed in section “[Experiments](#)”), a visual inspection pipeline was simulated (see Fig. 9). First, a stream of images is directed toward the machine learning model trained to identify possible defects. Then, based on the prediction score, a decision is made

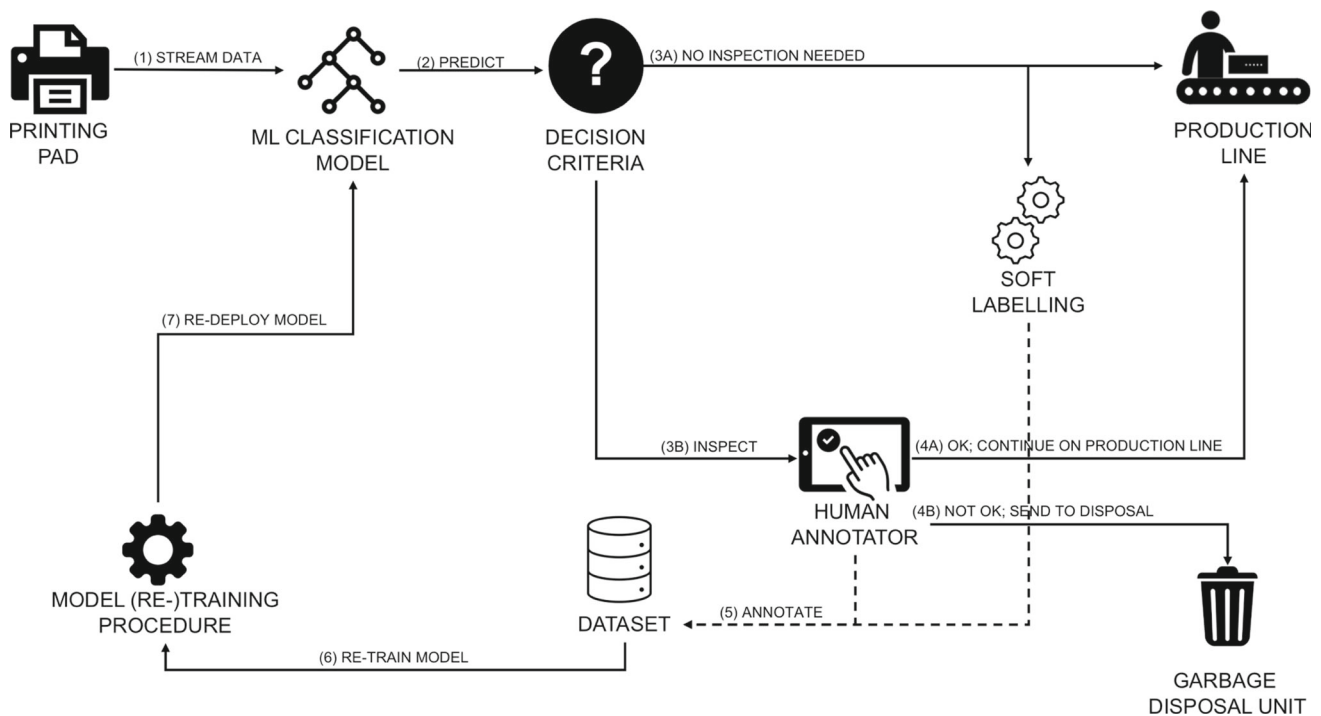
on whether the manufactured product should remain in the production line or be deferred to manual inspection. If the product is unlikely to be defective, such a decision can be considered a label (it is considered a soft label when not made by a human annotator). The label is then persisted, enlarging the existing dataset. The enlarged dataset can be used to retrain the model and replace the existing one after a successful deployment.

### Methodological aspects to evaluate probability calibration metrics and strategies

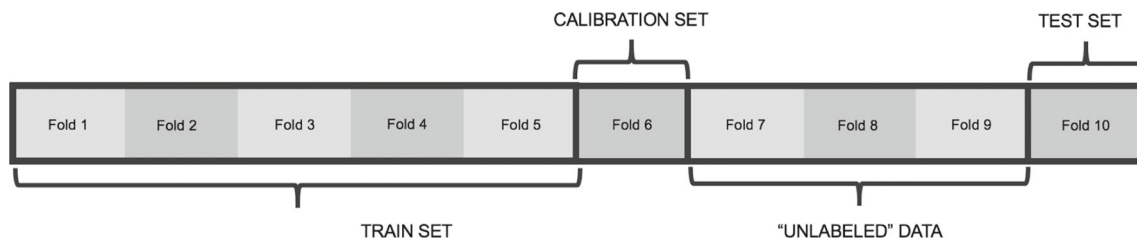
wECE metric definition.  $n$  indicates the number of classes under consideration.

$$wECE = \sum_{i=1}^n ECE_i \cdot w_i \quad (10)$$

A similar procedure was followed in the previous subsection to evaluate the proposed probability calibration metrics and techniques, avoiding the active learning step. Furthermore, a different dataset split was considered (see Fig. 10). After training the machine learning model and calibrating it with the calibration set, the non-calibrated model was used to issue a prediction for each instance in the *unlabeled data set*. The predicted class is then used to adjust further (train) the calibrator. While this introduces some noise, we expect that the better the classification model, the more it would benefit the calibrator, as explained in section “[Approximate model's probabilities calibration](#)”. Eleven performance metrics were measured: AUC ROC, ECE (computed as the ECE for each class and averaged, assigning the same weight to all classes),



**Fig. 9** Expected visual inspection pipeline in a production setting. Multiple active learning strategies were assessed to identify which would drive the best results



**Fig. 10** A 10-fold stratified cross-validation was used. The data was split into train set, test set, calibration set, and *unlabeled data*. *Unlabeled data* was used to simulate a stream of unlabeled data and assess whether a histogram-based calibration method without ground truth can enhance its performance over time

wECE (computed as the classwise ECE—see Eq. 10), PCS, wPCS,  $APCS_{\mathbb{W}}$ ,  $wAPCS_{\mathbb{W}}$ , APCS, wAPCS, MPCs, and wMPCS. AUC ROC measures the discriminative capability of the model and provides insights into how such capability is affected by different calibration techniques. ECE evaluates the expected difference between the accuracy and confidence of a calibration model. The ECE metric was used to compare the calibration quality for the multiple calibration techniques and the newly proposed PCS, wPCS, APCS, wAPCS, MPCs, and wMPCS metrics. Furthermore, given that the newly proposed metrics were built on a similar concept as the ECE metric, we are interested in how much they capture the same information. The Kendall  $\tau$  [see Kendall (1938)] and the Pearson correlation between ECE and the newly proposed metrics were measured. The Kendall correlation measures the ordinal association between two measured quantities. In this case, it measures to what extent both metrics increase or

decrease, given the predictions for a given machine learning model and calibrators. The Pearson correlation, on the other side, was used to assess whether the correlation between metrics was linear.

The metrics were computed on the test set against the ground truth (class annotations) and the approximate ground truth (predicted classes). The results were analyzed to understand how well the metrics capture the models' performance and calibration when no ground truth is available. Furthermore, the weighted and non-weighted metrics were compared to understand how class weighting influences the final score and perception regarding the quality of the calibration.

## Experiments

### Experimenting with active learning strategies

For this research, two active learning settings were explored (pool-based and stream-based), using four distinct strategies to label the queried data instances in an active learning setting. Two strategies were used to select data from the active learning set under the pool-based active learning setting: (a) random sampling and (b) instances for which the classification model was the most uncertain. The model's uncertainty was assessed by considering the highest score for a given class for a given instance and selecting the instance with the lowest score among the scores provided for the data instances in the active learning set. In both cases, data were sampled until the set's exhaustion. Under the streaming active learning setting, a slightly different policy was used. When random sampling was used, a decision was made whether to keep or discard the instance with a probability threshold of 0.5. Under the highest uncertainty selection criteria, the prediction for each data instance was analyzed and derived to the oracles for labeling if it was below a certain confidence threshold ( $p = 0.95$  or  $p = 0.99$ ).

Three oracle settings were considered (see Fig. 11): (A) human labeler as the only source of truth, (B) machine oracle (classifier model) for data instances where the classifier had a high certainty, and a human labeler otherwise; and (C) machine oracle (classifier model) for data instances where the classifier had a high certainty, and requesting an additional *opinion* to another machine oracle when uncertain about the outcome. This second oracle queries the closest labeled image from three randomly picked images (one per class). In (C), the machine oracle issues a label only when both machine oracles are unanimous on the label; otherwise, the instance labeling is delegated to a human labeler. The decision regarding which oracle to query was based on the models' confidence regarding the outcome and a probability threshold set based on manufacturing quality policies. It was assumed that the second machine oracle in (C) is accessible at a certain cost (e.g., paid external service) and, therefore, cannot be used for every prediction. Such a service was simulated by computing the Structural Similarity Index Measure (SSIM) score over the queried image.

Eight scenarios were set up (see Table 1), and experimented with two quality thresholds (0.95 and 0.99 probability that the item corresponded to a certain class) and five machine learning models. The machine learning models were calibrated using a sigmoid model based on Platt logistic model (Platt, 1999) (see Eq. 11).

Platt classifier calibration logistic model.  $y_i$  denotes the truth label, and  $f_i$  denotes the uncalibrated classifier's prediction for a particular sample.  $A$  and  $B$  denote adjusted

parameters when fitting the regressor.

$$P(y_i = 1 | f_i) = \frac{1}{1 + \exp(Af_i + B)} \quad (11)$$

### Experiments assessing probability calibration metrics and techniques

In an automated visual inspection setting, a labeling effort is required to (a) label data to train and calibrate the machine learning models and (b) perform a manual inspection when the models cannot determine the class of a given data instance accurately. To understand how the probability calibration affects the machine learning models, the models' predictions were compared against those obtained by (a) not calibrating the model (*No calibration*) and calibrating the model with (b) a sigmoid model based on the Platt logistic model (*Platt*), (c) temperature scaling (*Temperature*), and (d) *Histogram* calibration. Two aspects were considered in the experiments: (i) how calibration techniques compare against each other and (ii) whether calibrating a model without a ground truth can provide comparable results to models calibrated with ground truth.

## Results and evaluation

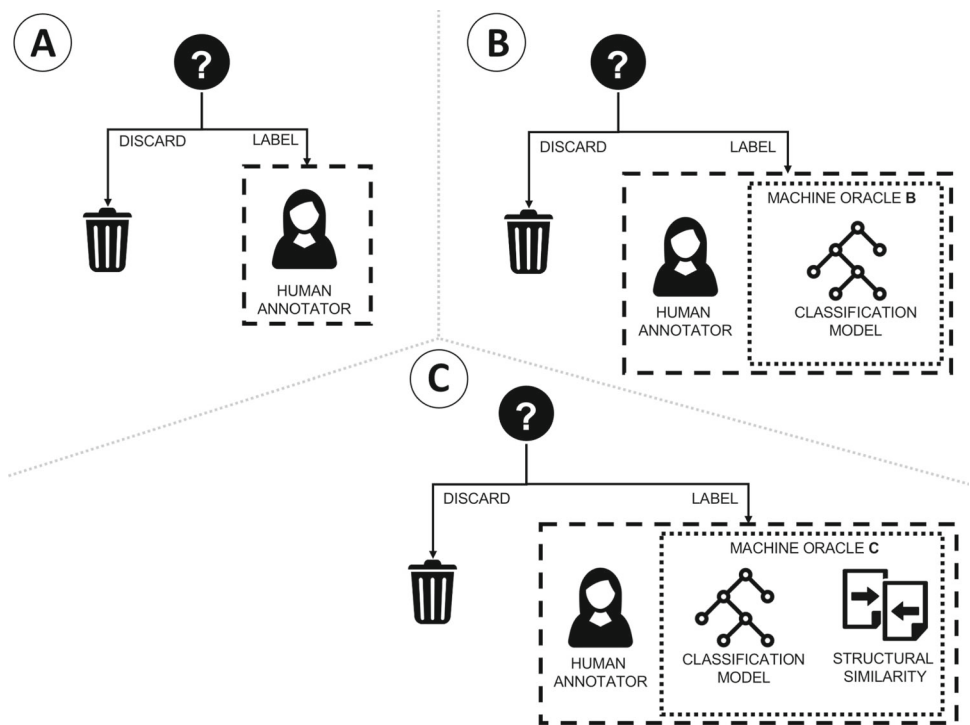
### Results and evaluation of active learning strategies

The active learning strategies were analyzed from two points of view. First, whether they contributed to better learning of the machine learning model. Second, how much manual work could be saved by adopting such strategies.

For the first case, the AUC ROC was measured over time (see Table 2). In particular, the models' average performance was contrasted when they consumed data within the Q1 and Q4 of the active learning pool. The best outcomes were observed for Experiment 2 (highest uncertainty with human labeler) settings, while the second-best performance was observed for Experiment 8 (highest uncertainty, with the machine and human oracles). Overall, it was observed that the streaming setting had a better average performance when compared to the pool-based experiments, despite achieving only the second-best results with Platt scaling. Furthermore, in two cases, the machine learning model degraded its performance between Q1 and Q4. This happened for Experiment 3 ( $p = 0.95$ ) and Experiment 4 ( $p = 0.95$ ).

Given that (a) in both experiments, a machine oracle was used, (b) no performance decrease was observed for  $p = 0.99$ , and (c) that the same setting did not affect the streaming case, we were tempted to conclude that most likely the machine oracles mislabeled certain instances, confusing the model when retrained and therefore reducing the

**Fig. 11** Three oracle settings are explored in this research: **A** human annotator, **B** soft-labeling with classification model’s outcomes for instances with high-confidence scores, and human annotator for instances where the model has low confidence; and **C** which is analogous to **B**, but the machine oracle takes into account the classifier’s output score and whether the predicted class matches the class with the shortest distance towards the active sample. In **C**, the sample is sent to manual revision if there is a class mismatch in the machine oracle. Samples are only discarded in a streaming setting



**Table 1** Proposed experiments to evaluate the best active learning setting regarding how it influences the models’ learning and its impact on the manual revision workload

Experiment	AL setting	AL data selection	Oracle
1	Pool-based	Random sampling	Human labeler
2	Pool-based	Highest uncertainty	Human labeler
3	Pool-based	Highest uncertainty	Machine Oracle B + Human labeler
4	Pool-based	Highest uncertainty	Machine Oracle C + Human labeler
5	Stream-based	Random sampling	Human labeler
6	Stream-based	Highest uncertainty	Human labeler
7	Stream-based	Highest uncertainty	Machine Oracle B + Human labeler
8	Stream-based	Highest uncertainty	Machine Oracle C + Human labeler

**Table 2** Mean values for the mean AUC ROC computed across ten folds for five machine learning models

Setup	Experiment	p=0.95		p=0.99	
		Q1	Q4	Q1	Q4
Pool-based	1-Random, human oracle	0.8428	0.8612	0.8431	0.8623
	2-Uncertainty, human oracle	<b>0.8594</b>	<b>0.8693</b>	<b>0.8594</b>	<b>0.8693</b>
	3-Uncertainty, oracle (machine B + human)	0.8398	0.8396	0.8398	0.8398
	4-Uncertainty, oracle (machine C + human)	0.8349	0.8348	0.8358	0.8358
Streaming	5-Random, human oracle	0.8460	0.8559	0.8460	0.8559
	6-Uncertainty, human oracle	0.8525	0.8647	0.8529	0.8647
	7-Uncertainty, oracle (machine B + human)	0.8505	0.8608	0.8529	0.8647
	8-Uncertainty, oracle (machine C + human)	<i>0.8550</i>	<i>0.8665</i>	<i>0.8553</i>	<i>0.8668</i>

The results show how different active learning policies influence the models’ learning over time [Q1 (first quartile) vs. Q4 (last quartile)]. Two probability thresholds (0.95 and 0.99) were considered as a soft labeling cut-off. The best results are bolded, and the second-best ones are displayed in italics

model's performance over time. Nevertheless, further analysis revealed a small fraction of soft-labeled data and that most cases were accurately labeled. While soft labeling was detrimental for the pool-based active learning settings, it led to superior results in a streaming setting, achieving results close to the best ones obtained across all experiments.

In Table 3 we report the performance of machine learning models for Experiment 2 and compare how they performed after Q1 and Q4 of the active learning pool data was shown to them. We found that the best performance was attained by the MLP, followed by the SVM by at least 0.05 AUC ROC points. Furthermore, while the MLP increased its performance over time, the SVM slightly reduced it in Q4. No other model had a performance decrease over time. Since Experiment 2 only considered a human oracle and the annotations are accurate, the performance decrease cannot be attributed to mislabeling. Furthermore, while the model's discriminative capacity loss could be attributed to the class imbalance, we consider this improbable, given that the rest of the models could better discern among the classes over time. Finally, the CART model obtained the worst results, which lagged slightly more than 0.16 AUC ROC points compared to the best one.

As mentioned at the beginning of this section, another relevant aspect of evaluating active learning strategies is their potential to reduce data annotation efforts. This could be analyzed from two perspectives. First, whether the additional data annotations provide enough knowledge to enhance the models' performance significantly. If not, the data annotation can be avoided. Second, a strategy can be devised (e.g., a machine oracle) to reduce the manual annotation effort. In this work, we focused on the second one. Table 4 presents the results for a cut-off value of  $p = 0.95$ . For  $p = 0.99$ , no instances were retrieved and given to machine oracles; therefore, no analysis was performed on them. The task required annotating 2460 samples on average.

When considering the cut-off value of 0.95, it was noticed that the Platt calibration considered a negligible number of cases for each experiment. While the quality of the annotations was high, using machine oracles would not strongly alleviate the manual labeling effort. The highest amount of soft-labeled instances corresponded to experiments with streaming settings (Experiment 7 and Experiment 8), which soft-labeled 4% and 3% of all data instances, respectively. Furthermore, 96% of samples were correctly labeled in both cases, meeting the quality threshold of  $p = 0.95$ . The decrease in the amount of soft labeled samples for Experiment 8 was due to discrepancies between the machine learning model and the SSIM score. Furthermore, the best machine labeling quality was achieved when considering *Oracle C* (unanimous vote of two machine oracles). When contrasting with the AUC ROC results obtained for those experiments, it was observed that while Experiments 3 and 4 slightly decreased

discriminative power, Experiments 7 and 8 increased their performance for at least 0.01 AUC ROC points.

## Results and evaluation of probability calibration metrics and techniques

The experiments performed in this research aimed to validate whether the metrics proposed to measure the quality of a calibrator can be used to understand the performance of a calibrator even when no ground truth is available. Furthermore, it aimed to validate whether predictions on unlabeled data could enhance the calibrators' performance. The results are presented in Tables 5, 6, and 7. The PCS, APCS, and MPCS (along with the weighted variants) metrics were computed considering the PPCM histogram, which denotes a perfect calibration.

To understand whether the proposed metrics can measure the calibration quality without ground truth, the Pearson and Kendall correlations were computed between the ECE, wECE, APCS<sub>W</sub>, wAPCS<sub>W</sub>, PCS, and wPCS metrics (see Table 5). While ECE and wECE are always computed considering the ground truth at the test set, PCS, APCS<sub>W</sub>, wPCS, and wAPCS<sub>W</sub> were calculated considering two cases: ground truth (golden standard) and predicted labels (approximate ground truth) at the test set. Furthermore, the correlations between the metrics were evaluated in two separate moments: after calibrating the models with the calibration set (CS) and after calibrating the models with additional samples retrieved from the *unlabeled data* set (CS+UD). The results show that the correlation between ECE, PCS, APCS<sub>W</sub>, wPCS, and wAPCS<sub>W</sub> metrics is consistent across all cases. Furthermore, little variation exists between the values obtained when PCS or wPCS were computed on the ground truth or the approximate ground truth. While the Pearson correlation decreases after training the calibrator with predicted labels from the unlabeled data set, the Kendall correlation grew stronger when PCS or APCS<sub>W</sub> were just averaged across classes and not weighted by the frequency of occurrence of each class. We consider the correlations moderate (negative Pearson correlation was measured between 0.50 and 0.61) or strong (negative Kendall correlation was above 0.33 and slightly below 0.40). Given the abovementioned results, we consider the PCS, wPCS, APCS<sub>W</sub>, and wAPCS<sub>W</sub> metrics adequately capture information conveyed by the ECE metric regardless of the source of truth used to measure the quality of the calibration. Therefore, we conclude that PCS, wPCS, APCS<sub>W</sub>, and wAPCS<sub>W</sub> can be used to assess the calibrators' quality when no ground truth is available.

Tables 6 and 7, compare the calibrators across multiple metrics to assess how an approximate calibration affects their discriminative power (AUC ROC) and whether it helps to enhance the calibrators' quality (ECE, PCS, APCS, MPCS,



**Table 3** Mean AUC ROC values computed across ten test folds for five machine learning models

Model	Q1	Q4	DS(p=0.95)
MLP	<b>0.9309±0.0004</b>	<b>0.9448±0.0003</b>	Yes
SVM	<i>0.8788±0.0007</i>	<i>0.8767±0.0007</i>	Yes
NB	0.8628±0.0005	0.8675±0.0005	Yes
KNN	0.8575±0.0006	0.8720±0.0006	Yes
CART	0.7669±0.0007	0.7854±0.0008	Yes

The results show how the machine learning models learn over time (Q1 vs. Q4) under the Experiment 2 setting. Furthermore, we analyze if the differences were statistically significant at a p-value = 0.95 [DS(p = 0.95)]. The best results are bolded, and the second-best results are displayed in italics

**Table 4** Proportion and quality of soft labeling through different settings, considering a predicted probability cut-off value of p = 0.95

Experiment	p=0.95 SL (%)	SL OK (%)	ML SL OK (%)	SSIM SL OK (%)
3	0.0077	0.9684	0.0075	NA
4	0.0033	0.9756	0.0050	0.0033
7	0.0413	0.9685	0.0400	NA
8	0.0343	0.9692	0.0483	0.0334

*SL (%)* denotes the percentage of soft annotated data instances w.r.t. the total, *SL OK (%)* denotes the percentage of correctly soft annotated instances, *ML SL OK (%)* denotes the percentage of soft annotated data instances w.r.t. the total that would be correctly annotated considering the ML model score, *SSIM SL OK (%)* denotes the percentage of soft annotated data instances w.r.t. the total that would be correctly annotated considering the SSIM score

**Table 5** The results were obtained for different models and probability calibration techniques

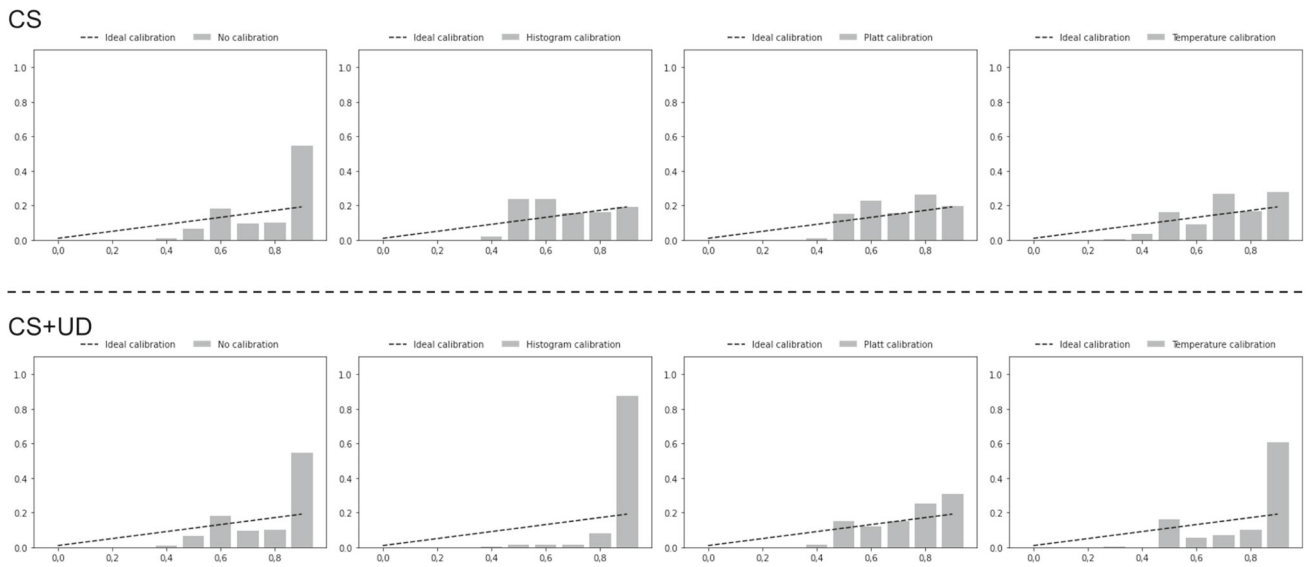
Source of truth	Correlation	Pearson		Kendall	
	Calibration data	CS	CS+UD	CS	CS+UD
Golden standard	ECE vs. PCS	− 0.6100	− 0.5937	− 0.3360	− 0.3981
	wECE vs. wPCS	− 0.5113	− 0.5070	− 0.3574	− 0.3525
	ECE vs. APCS <sub>W</sub>	− 0.6100	− 0.5939	− 0.3316	− 0.3981
	wECE vs. wAPCS <sub>W</sub>	− 0.5112	− 0.5070	0.3574	− 0.3480
Predicted labels	ECE vs. PCS	− 0.6100	− 0.5937	− 0.3360	− 0.3981
	wECE vs. wPCS	− 0.5084	− 0.5017	− 0.3360	− 0.3308
	ECE vs. APCS <sub>W</sub>	− 0.6100	− 0.5939	− 0.3316	− 0.3981
	wECE vs. wAPCS <sub>W</sub>	− 0.5083	− 0.5016	0.3423	− 0.3308

We show Pearson and Kendall correlation coefficients when comparing the ECE, APCS<sub>W</sub>, wAPCS<sub>W</sub>, PCS, and wPCS metrics. APCS<sub>W</sub>, wAPCS<sub>W</sub>, PCS, and wPCS are measured considering the ground truth (golden standard) or approximate ground truth (predicted labels). The ECE metric is always computed considering the ground truth. *CS* stands for *Calibration Set*, while *CS+UD* abbreviates *Calibration Set + Unlabeled Data*

and their weighted variants). Furthermore, Fig. 12 presents calibration plots for each calibrator for CS and CS+UD for visual assessment.

When comparing the calibrators through non-weighted metrics (see Table 6), we consider the Platt calibrator achieved the most stable performance. While the measured quality of calibration slightly decreased with the approximate calibration, it must be noticed that a higher proportion of positives was allocated at higher scores. Furthermore, while with the approximate calibration, the model's overall discriminative power slightly decreased, it remained superior against other models (even the not calibrated) by at least 0.02 AUC

ROC points. The Histogram and Temperature calibrators provide an interesting case, given both had a similar initial (CS) calibration quality if measured with the ECE, PCS, APCS, or MPCS metrics. Nevertheless, the metrics at CS+UD showed discrepancies: while ECE slightly increased for the Histogram calibrator (showing a worse calibration quality), it remained the same for the Temperature calibrator. On the other hand, PCS, APCS, and MPCS decreased (signaling a worse calibration quality) for both the Histogram and Temperature calibrator. Furthermore, the decrease in the metrics' values was more pronounced for the Histogram calibrator. When visually assessing both calibrators, we found that they



**Fig. 12** Eight calibration plots, comparing *No calibration*, *Histogram calibration*, *Platt calibration*, and *Temperature calibration* at CS (calibrated with the calibration set (ground truth)) and CS + UD (calibrated with a calibration set and predicted labels over time). The calibration plots have been adapted, showing normalized values (their sum is one)

rather than the usual fraction of positives on the dependent variable axis. The x-axis denotes the mean predicted probability for a given class. The histograms average predictions across classes and calibrated machine learning models

had a similar initial distribution (CS), but the Histogram calibrator ended up much more skewed than the Temperature calibrator at CS+UD. While the ECE metric did not capture this behavior, it was successfully summarized in the PCS, APCS, and MPCS metrics. We found the same patterns could be observed when analyzing the weighted metrics (see Table 7).

From the results above, we confirm that the proposed metrics can accurately measure the quality of calibration of a given calibrator when no ground truth is available. Furthermore, the metrics have shown to provide a more accurate measurement of the calibrators’ quality than ECE, overcoming some of its shortcomings (e.g., providing a more holistic

view of the distribution of positives along the mean predicted probability, taking into account empty bins).

Our research shows that tracking predictions over time did not enhance the quality of calibration for any of the methods involved (Histogram calibration, Platt calibration, or Temperature calibration). Finding accurate calibration models for probability calibration, given a lack of ground truth, remains a matter of future work.

**Table 6** The results were obtained for different probability calibration techniques

Source of truth	Calibration	AUC ROC ↑		ECE ↓		PCS ↑		APCS ↑		MPCS ↑	
		CS	CS+UD	CS	CS+UD	CS	CS+UD	CS	CS+UD	CS	CS+UD
Golden standard	None	0.8630	0.8630	0.1090	0.1090	0.7636	0.7636	0.7448	0.7448	0.5647	0.5647
	Histogram	0.8432	0.8442	0.1051	0.1084	0.8050	0.6509	0.7458	0.6697	0.5539	0.4507
	Platt	<b>0.8907</b>	0.8903	<b>0.0914</b>	0.0955	0.7523	0.7421	<b>0.7669</b>	0.7613	<b>0.5904</b>	0.5820
	Temperature	0.8614	0.8609	0.1090	0.1090	<b>0.8073</b>	0.7548	0.7651	0.7383	0.5886	0.5517
Predicted labels	None	0.8630	0.8630	0.1090	0.1090	0.7636	0.7636	0.7448	0.7448	0.5647	0.5647
	Histogram	0.8432	0.8442	0.1051	0.1084	0.8050	0.6509	0.7458	0.6697	0.5539	0.4507
	Platt	<b>0.8907</b>	0.8903	<b>0.0914</b>	0.0955	0.7523	0.7421	<b>0.7669</b>	0.7613	<b>0.5904</b>	0.5820
	Temperature	0.8614	0.8609	0.1090	0.1090	<b>0.8073</b>	0.7548	0.7651	0.7383	0.5886	0.5517

PCS, APCS, and MPCS are measured considering the ground truth (golden standard) or approximate ground truth (predicted labels). The AUC ROC and ECE metrics are always computed considering the ground truth. The best results are bolded

**Table 7** The results were obtained for different probability calibration techniques

Source of truth	Calibration	AUC ROC ↑		wECE ↓		wPCS ↑		wAPCS ↑		wMPCS ↑	
		CS	CS+UD	CS	CS+UD	CS	CS+UD	CS	CS+UD	CS	CS+UD
Golden standard	None	0.8630	0.8630	0.1457	0.1457	0.7829	0.7829	0.7545	0.7545	0.5801	0.5801
	Histogram	0.8432	0.8442	0.1410	0.1449	<b>0.8265</b>	0.6494	0.7565	0.6689	0.5685	0.4505
	Platt	<b>0.8907</b>	0.8903	<b>0.1230</b>	0.1285	0.7655	0.7527	<b>0.7735</b>	0.7666	0.6010	0.5902
	Temperature	0.8614	0.8609	0.1457	0.1457	0.8232	0.7773	0.7730	0.7496	<b>0.6014</b>	0.5697
Predicted labels	None	0.8630	0.8630	0.1457	0.1457	0.7826	0.7826	0.7543	0.7543	0.5799	0.5799
	Histogram	0.8432	0.8442	0.1410	0.1449	<b>0.8265</b>	0.6494	0.7565	0.6689	0.5686	0.4505
	Platt	<b>0.8907</b>	0.8903	<b>0.1230</b>	0.1285	0.7641	0.7517	0.7728	0.7662	0.5998	0.5894
	Temperature	0.8614	0.8609	0.1457	0.1457	0.8230	0.7765	0.7730	0.7492	0.6013	0.5692

wPCS, wAPCS, and wMPCS are measured considering the ground truth (golden standard) or approximate ground truth (predicted labels). The AUC ROC and wECE metrics are always computed considering the ground truth. The best results are bolded

## Conclusions and future work

This work explored active learning with multiple oracles to alleviate the manual inspection of manufactured products and the labeling of inspected products. Our active learning settings can save up to four percent of the manual inspection and data labeling load while not compromising on the quality of the outcome for a quality threshold of  $p = 0.95$ . It must be noted that labeling savings depend on the machine learning model deployed, the acceptance quality levels, and the quality of the active learning machine oracles under consideration. Furthermore, multiple probability calibration techniques were compared, and several new metrics to measure the quality of a calibrator were proposed. The metrics enable measuring the calibrators' quality even when no ground truth is available. The experiments demonstrated that the proposed metrics capture relevant data otherwise summarized in the ECE metric—a popular metric to measure the quality of a probability calibration model. Nevertheless, the behavior of the proposed metrics under concept drift was not studied yet, and we consider it a matter of future research.

We envision multiple lines of investigation for future work. Regarding active learning, we are interested in enriching our current setup by adopting different strategies to decide how interesting an upcoming image is (e.g., learning distance metrics for each class or learning to predict which piece of data would enhance the classifier the most) and enhancing the calibration techniques to display the desired behavior for high-confidence thresholds. We will conduct further research on probabilities calibration to understand how the proposed metrics behave when concept drift occurs. Finally, we will explore new approximate probability calibration approaches leading to enhanced calibrators when no ground truth is available.

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

**Author Contributions** JMR: conceptualization, methodology, software programming, validation, formal analysis, investigation, writing (original draft, review and editing), visualization, supervision. LB: conceptualization, validation, writing (review and editing). ET: software programming, writing (review and editing). PZ: software programming, writing (review and editing). JK: resources, data curation. BF: resources, writing (review and editing), supervision, project administration, funding acquisition. DM: resources, writing (review and editing), supervision, project administration, funding acquisition.

**Data availability** The datasets analyzed during the current study are not publicly available for confidentiality reasons.

## Declarations

**Conflict of interest** The authors have no competing interests to declare relevant to this article's content.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aggour, K. S., Gupta, V. K., Ruscitto, D., Ajdelsztajn, L., Bian, X., Brosnan, K. H., Kumar, N. C., Dheeradhada, V., Hanlon, T., Iyer, N., et al. (2019). Artificial intelligence/machine learning in manufacturing and inspection: A GE perspective. *MRS Bulletin*, 44(7), 545–558. <https://doi.org/10.1557/mrs.2019.157>
- Aiger, D., & Talbot, H. (2012). The phase only transform for unsupervised surface defect detection. In S. Barthel, J. Colding, & T. Elmqvist (Eds.), *Emerging topics in computer vision and its appli-*

- ations (pp. 215–232). World Scientific. <https://doi.org/10.1109/CVPR.2010.5540198>
- Aminzadeh, M., & Kurfess, T. R. (2019). Online quality inspection using Bayesian classification in powder-bed additive manufacturing from high-resolution visual camera images. *Journal of Intelligent Manufacturing*, 30, 2505–2523.
- Barari, A., de Sales Guerra Tsuzuki, M., Cohen, Y., & Macchi, M. (2021). Intelligent manufacturing systems towards industry 4.0 era. *Journal of Intelligent Manufacturing*, 32, 1793–1796.
- Beltrán-González, C., Bustreo, M., & Del Bue, A. (2020). External and internal quality inspection of aerospace components. In *2020 IEEE 7th international workshop on metrology for aerospace (MetroAeroSpace)*, (pp. 351–355). IEEE. <https://doi.org/10.1109/MetroAeroSpace48742.2020.9160103>
- Beluch, W. H., Genewein, T., Nürnberger, A., & Köhler, J. M. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 9368–9377). <https://doi.org/10.1109/CVPR.2018.00976>
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
- Bröcker, J., & Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3), 651–661. <https://doi.org/10.1175/WAF993.1>
- Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). *API design for machine learning software: experiences from the scikit-learn project* (pp. 108–122). In ECML PKDD Workshop: Languages for Data Mining and Machine Learning.
- Carvajal Soto, J., Tavakolizadeh, F., & 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. <https://doi.org/10.1080/0951192X.2019.1571238>
- Cheeseman, P. C. (1985). In defense of probability. In *IJCAI*, 85, 1002–1009.
- Chouchene, A., Carvalho, A., Lima, T. M., Charrua-Santos, F., Osório, G. J., & 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)*, (pp. 127–131). IEEE. <https://doi.org/10.1109/ICITM48982.2020.9080396>
- Cohen, I. & Goldszmidt, M. (2004). Properties and benefits of calibrated classifiers. In *European conference on principles of data mining and knowledge discovery*, (pp. 125–136). Springer. [https://doi.org/10.1007/978-3-540-30116-5\\_14](https://doi.org/10.1007/978-3-540-30116-5_14)
- Cohen, N. & Hoshen, Y. (2020). Sub-image anomaly detection with deep pyramid correspondences. [arXiv:2005.02357](https://arxiv.org/abs/2005.02357)
- Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15(2), 201–221. <https://doi.org/10.1007/BF00993277>
- Cordier, A., Das, D., & Gutierrez, P. (2021). Active learning using weakly supervised signals for quality inspection. [arXiv:2104.02973](https://arxiv.org/abs/2104.02973)
- Cullinane, S.-J., Bosak, J., Flood, P. C., & Demerouti, E. (2013). Job design under lean manufacturing and its impact on employee outcomes. *Organizational Psychology Review*, 3(1), 41–61. <https://doi.org/10.1177/2041386612456412>
- Dai, W., Mujeeb, A., Erdt, M., & Sourin, A. (2018). Towards automatic optical inspection of soldering defects. In *2018 International Conference on Cyberworlds (CW)*, (pp. 375–382). IEEE. <https://doi.org/10.1109/CW.2018.00074>
- Duan, G., Wang, H., Liu, Z., & Chen, Y.-W. (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), 1679–1689. <https://doi.org/10.1109/TSMCC.2012.2216260>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, (pp. 1321–1330). PMLR.
- Gupta, C. & Ramdas, A. (2021). Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, (pp. 3942–3952). PMLR.
- 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*, (pp. 770–778). <https://doi.org/10.48550/arXiv.1512.03385>
- Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. R. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, 21(8), 1509–1515. <https://doi.org/10.1093/bioinformatics/bti171>
- Jezek, S., Jonak, M., Burget, R., Dvorak, P., & Skotak, M. (2021). Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, (pp. 66–71). IEEE. <https://doi.org/10.1109/ICUMT54235.2021.9631567>
- Jiang, J. & Wong, W. (2018). Fundamentals of common computer vision techniques for textile quality control. In *Applications of computer vision in fashion and textiles*, (pp. 3–15). Elsevier. <https://doi.org/10.1016/B978-0-08-101217-8.00001-4>
- Jian, C., Gao, J., & Ao, Y. (2017). Automatic surface defect detection for mobile phone screen glass based on machine vision. *Applied Soft Computing*, 52, 348–358. <https://doi.org/10.1016/j.asoc.2016.10.030>
- Kang, G.-W. & Liu, H.-B. (2005). Surface defects inspection of cold rolled strips based on neural network. In *2005 international conference on machine learning and cybernetics*, (Vol. 8, pp. 5034–5037). IEEE. <https://doi.org/10.1109/ICMLC.2005.1527830>
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2), 81–93. <https://doi.org/10.2307/2332226>
- Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling* (Vol. 26). Springer. <https://doi.org/10.1007/978-1-4614-6849-3>
- Kujawińska, A., Vogt, K., & Hamrol, A. (2016). The role of human motivation in quality inspection of production processes. In *Advances in ergonomics of manufacturing: managing the enterprise of the future*, (pp. 569–579). Springer. <https://doi.org/10.1007/978-1-4614-6849-3>
- Kumar, A., Liang, P. S., & Ma, T. (2019). Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 1–8.
- Küppers, F., Kronenberger, J., Shantia, A., & Haselhoff, A. (2020). Multivariate confidence calibration for object detection. In *The IEEE/CVF conference on computer vision and pattern recognition (CVPR) Workshops*
- Kurniati, N., Yeh, R.-H., & Lin, J.-J. (2015). Quality inspection and maintenance: The framework of interaction. *Procedia Manufacturing*, 4, 244–251. <https://doi.org/10.1016/j.promfg.2015.11.038>
- Leathart, T., Frank, E., Holmes, G., & Pfahringer, B. (2017). Probability calibration trees. In *Asian Conference on Machine Learning*, (pp. 145–160). PMLR
- Lewis, D. D. & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, (pp. 148–156). Elsevier. <https://doi.org/10.1016/B978-1-55860-335-6.50026-X>
- Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. (2021). Cutpaste: Self-supervised learning for anomaly detection and localization. In

- Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 9664–9674)
- Lin, H., Li, B., Wang, X., Shu, Y., & Niu, S. (2019). Automated defect inspection of led chip using deep convolutional neural network. *Journal of Intelligent Manufacturing*, 30, 2525–2534.
- Meng, L., McWilliams, B., Jarosinski, W., Park, H.-Y., Jung, Y.-G., Lee, J., & Zhang, J. (2020). Machine learning in additive manufacturing: A review. *JOM Journal of the Minerals Metals and Materials Society*, 72(6), 2363–2377. <https://doi.org/10.1007/s11837-020-04155-y>
- Mujeeb, A., Dai, W., Erdt, M., & Sourin, A. (2018). Unsupervised surface defect detection using deep autoencoders and data augmentation. In *2018 International conference on cyberworlds (CW)*, (pp. 391–398). IEEE. <https://doi.org/10.1109/CW.2018.00076>
- Newman, T. S., & Jain, A. K. (1995). A survey of automated visual inspection. *Computer Vision and Image Understanding*, 61(2), 231–262. <https://doi.org/10.1006/cviu.1995.1017>
- Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. In *CVPR Workshops*, volume 2.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 1–12.
- Park, J.-K., Kwon, B.-K., Park, J.-H., & Kang, D.-J. (2016). Machine learning-based imaging system for surface defect inspection. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 3(3), 303–310. <https://doi.org/10.1007/s40684-016-0039-x>
- Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6), 355–607.
- Platt, J., et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 61–74.
- Platt, J. C. (2000). *5 Probabilities for SV machines*. *Advances in large margin classifiers* (p. 61). MIT Press.
- Posocco, N. and Bonnefoy, A. (2021). Estimating expected calibration errors. In *International conference on artificial neural networks*, (pp. 139–150). Springer. <https://doi.org/10.1007/978-3-030-86380-712>
- Rai, R., Tiwari, M. K., Ivanov, D., & Dolgui, A. (2021). Machine learning in manufacturing and industry 4.0 applications. <https://doi.org/10.1080/00207543.2021.1956675>
- Ren, R., Hung, T., & Tan, K. C. (2017). A generic deep-learning-based approach for automated surface inspection. *IEEE Transactions on Cybernetics*, 48(3), 929–940. <https://doi.org/10.1109/TCYB.2017.2668395>
- Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., & Wang, X. (2020). A survey of deep active learning. [arXiv:2009.00236](https://arxiv.org/abs/2009.00236)
- Rippel, O., Haumering, P., Brauers, J., & Merhof, D. (2021). Anomaly detection for the automated visual inspection of pet preform closures. In *2021 26th IEEE international conference on emerging technologies and factory automation (ETFA)*, (pp. 1–7). IEEE. <https://doi.org/10.1109/ETFA45728.2021.9613298>
- Rožanec, J. M., Novalija, I., Zajec, P., Kenda, K., Tavakoli Ghinani, H., Suh, S., Veliou, E., Papamartzivanos, D., Giannetos, T., Menesidou, S.-A., et al. (2022). Human-centric artificial intelligence architecture for industry 5.0 applications. *International Journal Production Research*. <https://doi.org/10.1080/00207543.2022.2138611>
- Rožanec, J. M., Trajkova, E., Dam, P., Fortuna, B., & Mladenici, D. (2022). Streaming machine learning and online active learning for automated visual inspection. *IFAC-PapersOnLine*, 55(2), 277–282. <https://doi.org/10.1016/j.ifacol.2022.04.206>
- Schmitt, J., Bönig, J., Borggräfe, T., Beiting, G., & Deuse, J. (2020). Predictive model-based quality inspection using machine learning and edge cloud computing. *Advanced Engineering Informatics*, 45, 101101. <https://doi.org/10.1016/j.aei.2020.101101>
- See, J. E. (2012). Visual inspection: A review of the literature. Sandia Report SAND2012-8590, Sandia National Laboratories, Albuquerque, New Mexico. <https://doi.org/10.2172/1055636>
- Selvi, S. S. T., & Nasira, G. (2017). An effective automatic fabric defect detection system using digital image processing. *Journal of Water and Environmental Nanotechnology*, 6(1), 79–85. <https://doi.org/10.13074/jent.2017.03.171241>
- Settles, B. (2009). Active learning literature survey.
- Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2021). Classifier calibration: How to assess and improve predicted class probabilities: A survey. (pp. arXiv–2112). <https://doi.org/10.48550/arXiv.2112.10327>
- Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P., et al. (2021). Classifier calibration: How to assess and improve predicted class probabilities: A survey. [arXiv:2112.10327](https://arxiv.org/abs/2112.10327)
- Tsai, D.-M., & Lai, S.-C. (2008). Defect detection in periodically patterned surfaces using independent component analysis. *Pattern Recognition*, 41(9), 2812–2832. <https://doi.org/10.1016/j.patcog.2008.02.011>
- Valavanis, I., & Kosmopoulos, D. (2010). Multiclass defect detection and classification in weld radiographic images using geometric and texture features. *Expert Systems with Applications*, 37(12), 7606–7614. <https://doi.org/10.1016/j.eswa.2010.04.082>
- van Garderen, K. (2018). Active learning for overlay prediction in semiconductor manufacturing.
- Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1), 175–186. <https://doi.org/10.1007/s00521-013-1368-0>
- Vergara-Villegas, O. O., Cruz-Sánchez, V. G., Jesús Ochoa-Domínguez, H. d., Jesús Nandayapa-Alfaro, M. d., & Flores-Abad, Á. (2014). Automatic product quality inspection using computer vision systems. In *Lean manufacturing in the developing World*, (pp. 135–156). Springer.
- Villalba-Diez, J., Schmidt, D., Gevers, R., Ordieres-Meré, J., Buchwitz, M., & Wellbrock, W. (2019). Deep learning for industrial computer vision quality control in the printing industry 4.0. *Sensors*, 19(18), 3987. <https://doi.org/10.3390/s19183987>
- Villani, C. (2009). *Optimal transport: Old and new* (Vol. 338). Springer.
- Weiss, S. M., Dhurandhar, A., Baseman, R. J., White, B. F., Logan, R., Winslow, J. K., & Poindexter, D. (2016). Continuous prediction of manufacturing performance throughout the production lifecycle. *Journal of Intelligent Manufacturing*, 27(4), 751–763. <https://doi.org/10.1007/s10845-014-0911-x>
- Wuest, T., Irgens, C., & 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. <https://doi.org/10.1007/s10845-013-0761-y>
- Yang, J., Li, S., Wang, Z., Dong, H., Wang, J., & Tang, S. (2020). Using deep learning to detect defects in manufacturing: A comprehensive survey and current challenges. *Materials*, 13(24), 5755. <https://doi.org/10.3390/ma13245755>
- Yun, J. P., Choi, D.-C., Jeon, Y.-J., Park, C., & Kim, S. W. (2014). Defect inspection system for steel wire rods produced by hot rolling process. *The International Journal of Advanced Manufacturing Technology*, 70(9), 1625–1634. <https://doi.org/10.1007/s00170-013-5397-8>
- Zajec, P., Rožanec, J. M., Novalija, I., Fortuna, B., Mladenici, D., & Kenda, K. (2021). Towards active learning based smart assistant for manufacturing. In *IFIP international conference on advances in production management systems*, (pp. 295–302). Springer. <https://doi.org/10.1007/978-3-030-85910-731>

- Zavrtanik, V., Kristan, M., & Skočaj, D. (2021). Draem—a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, (pp. 8330–8339)
- Zavrtanik, V., Kristan, M., & Skočaj, D. (2022). Dsr—A dual subspace re-projection network for surface anomaly detection supplementary material.
- Zavrtanik, V., Kristan, M., & Skočaj, D. (2022). Dsr—A dual subspace re-projection network for surface anomaly detection. In *European conference on computer vision*, (pp. 539–554). Springer. <https://doi.org/10.1007/978-3-031-19821-231>
- Zeng, X., & Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1), 1–12. <https://doi.org/10.1080/095281300146272>
- Zheng, T., Ardolino, M., Bacchetti, A., & Perona, M. (2021). The applications of industry 4.0 technologies in manufacturing context: A systematic literature review. *International Journal of Production Research*, 59(6), 1922–1954. <https://doi.org/10.1080/00207543.2020.1824085>
- Zheng, Z., Zhang, S., Yu, B., Li, Q., & Zhang, Y. (2020). Defect inspection in tire radiographic image using concise semantic segmentation. *IEEE Access*, 8, 112674–112687. <https://doi.org/10.1109/ACCESS.2020.3003089>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.