Active Learning and Approximate Model Calibration for Automated Visual Inspection in Manufacturing

Jože M. Rožanec\textsuperscript{1,2,5*}, Luka Bizjak\textsuperscript{2}, Elena Trajkova\textsuperscript{3}, Patrik Zajec\textsuperscript{2}, Jelle Keizer\textsuperscript{4}, Blaž Fortuna\textsuperscript{5} and Dunja Mladenič\textsuperscript{2}

\textsuperscript{1}Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, 1000, Slovenia.
\textsuperscript{2}Jožef Stefan Institute, Jamova 39, Ljubljana, 1000, Slovenia.
\textsuperscript{3}Faculty of Electrical Engineering, University of Ljubljana, Tržaška c. 25, Ljubljana, 1000, Slovenia.
\textsuperscript{4}Philips Consumer Lifestyle BV, Oliemolenstraat 5, Drachten, The Netherlands.
\textsuperscript{5}Qlector d.o.o., Rovšnikova 7, Ljubljana, 1000, Slovenia.

*Corresponding author(s). E-mail(s): joze.rozanec@ijs.si; Contributing authors: luka.bizjak@ijs.si; trajkova.elena.00@gmail.com; patrik.zajec@ijs.si; jelle.keizer@philips.com; blaz.fortuna@qlector.com; dunja.mladenic@ijs.si;

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 (with single and multiple oracles) to visual inspection. We propose a novel approach to probabilities calibration of classification models and two new metrics to assess the performance of the calibration without the need for ground truth. We performed experiments on real-world data provided by Philips Consumer Lifestyle BV. Our results show that 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, we show that the proposed 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

1 Introduction

Quality control is a key part of the manufacturing process, which comprehends inspection, testing, and identification to ensure the manufactured products comply with specific standards and specifications [1–3]. For example, the inspection tasks aim to determine whether a specific part features assembly integrity, surface finish, and adequate geometric dimensions [4]. 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.

One of such quality inspection activities is the visual inspection, considered a bottleneck activity in some instances [5]. Visual inspection is associated with many challenges. Some visual inspections are considered to require a substantial amount of reasoning capability, visual abilities, and specialization [4]. 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) [6, 7]. 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, wellbeing, and motivation, and the managements’ support and communication [8–10]. 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 [11, 12]), and trends in manufacturing (e.g., the Industry 4.0 and Industry 5.0 paradigms [13]) 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 [14, 15]. Automated visual inspection was applied in several use cases in the past [16–19]. Nevertheless, it is considered that the
field is still in its early stages and that artificial intelligence has the potential to revolutionize product inspection [20].

While machine learning models can be trained to determine whether a manufactured piece is defective or not 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 [21], and (c) how to alleviate the manual labeling effort. We use active learning to enhance the classification model to address the first challenge. We compare pool-based and stream-based settings with different active learning sample query strategies across five machine learning algorithms. We use the Platt scaling, a popular probability calibration technique to address the second challenge. Finally, when addressing the reduction of manual labeling effort, we consider two scenarios: (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. We address (i) by exploring the usage of multiple oracles and soft labeling to reduce the manual inspection effort. Finally, to address (ii), we propose a variation of the histogram-based calibration that does not require a ground truth and a novel metric to measure the fidelity of such calibration to the expected ground truth.

In this work, we extend our previous research described in paper Streaming Machine Learning and Online Active Learning for Automated Visual Inspection [22]. While in that paper, we researched the impact of active learning on streaming algorithms, in this paper, we explore batch and online settings along with different active learning policies and oracles. Furthermore, we calibrate our models so that they issue probability calibrations. Finally, we develop a histogram-based probability calibration technique that requires no ground truth and a novel metric to measure probability calibration techniques’ performance. We performed our research 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.

We evaluate the discriminative capability of the classification models using the Area Under the Receiver Operating Characteristic Curve (AUC ROC, see [23]). 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 we evaluate the models in a multiclass setting, we compute the AUC ROC with a one-vs-rest strategy. Furthermore, we compare the performance of multiple probability calibration approaches through the Estimated Calibration Error (ECE) and two novel metrics we propose.

This paper is organized as follows. Section 2 describes the current state of the art and related works. Section 3 describes a novel method for probabilities calibration and two metrics we propose, while Section 4 describes the use case.
The main novelty regarding the proposed probabilities calibration technique and metrics is the ability to calibrate and measure calibration quality without needing ground truth data. Next, Section 5 provides a detailed description of the methodology we followed. We describe the experiments we performed in Section 6, and discuss the results obtained in Section 7. Finally, Section 8 presents our conclusions and outlines future work.

2 Related Work

In this section, we provide a short overview of three topics relevant to our research: (i) use of machine learning for quality inspection, (ii) active learning, and (iii) probabilities calibration. To each of them, we devote a brief subsection.

2.1 Machine Learning for Quality Inspection

A comprehensive and reliable quality inspection is often indispensable to the manufacturing process, and high inspection volumes turn inspection processes into bottlenecks [24]. 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 to early drop a product that will not meet quality expectations and therefore 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 [25]. Automated visual inspection refers to image processing techniques for quality control, usually applied in the production line of manufacturing industries [19]. It has been successfully applied to determine the end quality of the products and 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) [26]. 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 [27].

[28] considers three approaches 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). [29] describe how TFT-LCD panels and LCD color filters were inspected by comparing surface segments containing complex periodic patterns. [30] successfully applied feedforward networks to detect surface defects on cold-rolled strips. In the same line, [31] proposed a novel defect detection algorithm for steel wire rods produced by the hot rolling process. [32] compared multiple machine learning models (Support Vector Machine, Neural Network, and K-nearest neighbors
Active Learning and Approximate Model Calibration in Manufacturing

(\text{kNN}) on defect detection in weld images. [26] developed a Convolutional Neural Network (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. Multiple authors developed machine learning algorithms for visual inspection leveraging feature extraction from pre-trained models [33–35]. 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) [36], or employed autoencoders to find how a reference image differs from the expected pattern [37, 38].

2.2 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, we find (i) uncertainty sampling (select an unlabeled sample with the highest uncertainty, given a certain metric or machine-learning model [39]), or (ii) query-by-committee (retrieve the unlabeled sample with the highest disagreement between a set of forecasting models (committee) [40, 41]). 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 [42]. In addition, it has been demonstrated that ensemble-based active learning can effectively counteract class imbalance through new labeled image acquisition [43]. 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 [44].

Active learning was successfully applied in manufacturing, but scientific literature remains scarce on this domain [45]. Some use cases include the automatic optical inspection of printed circuit boards [46], media news recommendation in a demand forecasting setting [47], and the identification of the local displacement between two layers on a chip in the semiconductor industry [48].

2.3 Probabilities calibration

Probabilities are defined as a generalization of predicate calculus where given a truth value of a formula given the evidence (degree of plausibility) is generalized to a real number between zero and one [49]. 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. Among the advantages of the probability calibration, we find that providing reliable estimates of the true probability that a sample is a member of a class of interest (a) usually does not decrease the classification accuracy, (b) the calibration provides thresholds on the decision rules and therefore minimizes the classification error, (c) 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 [50, 51].

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. A different criterion is introduced with the concept of confidence calibration, which only aims to calibrate the classifier’s most likely predicted class [51].

Multiple probability calibration methods have been proposed in the scientific literature. Among them, we find the post-hoc techniques, which 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 [52]. 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 [53] 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 [54]. 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 [55]. 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 [56]. 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 [51].

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 [57]. On the other hand, validity plots aim to convey the bin frequencies for every bin and therefore provide valuable information regarding miscalibration bounds [58]. Among the metrics, we find the binary ECE, which 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. Finally, 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. While the ECE metric is widely accepted, research has shown is subject to shortcomings (e.g., using fixed calibration ranges some bins contain most of the data, losing sharpness), and there is agreement that how to measure the probabilistic calibration remains a challenge [59].

While many probability calibration methods and metrics have been developed, we found 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. Furthermore, it lacks mechanisms to adapt to concept drift. To address this void, we propose performing empirical binning regardless of the ground truth labels, assuming the classifier could perform with perfect discriminative power in the best case. Furthermore, we developed a metric to assess the discrepancy between such an ideal scenario, the classifier at hand, and the quality of probability calibration we achieved. By doing so, we can leverage a virtually infinite number of data instances to map the prediction scores to a set of bins and maintain an up-to-date probability calibration map. Such a characteristic enables perfect fidelity towards the prediction scores and enhances the probability calibration map over time if no concept drift exists. While such a schema can also provide robustness against concept drift remains a matter of future work.
3 Approximate Model’s Probabilities Calibration

Fig. 1: The figure illustrates two sample histograms: the histogram on the left corresponds to some sub-optimally calibrated classifier, where scores are distributed across multiple values. In contrast, the histogram on the right (reference histogram) corresponds to a perfectly calibrated classifier. The Wasserstein distance between both histograms is the minimum distance between the existing calibration and a perfect one. The distance is weighted with the improvement opportunity regarding the specific classification model.

This research proposes an approximate approach to calibrating machine learning prediction scores to probabilities. The approach is based on histogram calibration, building an initial histogram based on a calibration data set, as it is common practice for probability calibration methods. In a calibration set, we have (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 for a histogram-based calibration.

We consider the final prediction of a calibrated model has 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 metrics and plots exist to assess the quality of the probability calibration, such means require a ground truth on which 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 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.

To address the abovementioned issues, we propose performing an Approximate Histogram-based Probability Calibration (AHPC). Under the assumption of a perfect classifier, the prediction scores should perfectly discriminate between classes. Therefore, if each prediction equals the ground truth, the calibration technique must only ensure a mapping between the predictive scores and the probabilistic outcome so that the mean predicted probability corresponds to the likelihood of occurrence of positive class events for a given class, and no additional information is required to calibrate the model. Such mapping can be achieved with a histogram, counting the occurrences of the prediction scores with one histogram per class. Furthermore, to ensure we can process a stream with infinite values, we instantiate the histograms with the \textit{t-digest} \cite{prata2017t} online algorithm. The \textit{t-digest} builds small sketches of data to approximate rank-based statistics with high accuracy while remaining robust to skewed distributions. We normalize such predictions to be contained between zero and one and determine the same number of fixed-width bins across classes to ensure the histograms remain comparable.

Given that no classifier performs perfectly, we expect a histogram of predicted probabilities to adopt different shapes. Nevertheless, we know the histogram shape of a perfectly calibrated model for a given class (see Fig. 1). We exploit this fact to propose a metric to measure the goodness of calibration while considering the imperfection of the classification model. To estimate how close the AHPC (or any other probability calibration method) is w.r.t. to the target (ideal) histogram, we use optimal transport \cite{rubner1998survey, cuturi2013sinkhorn}, in particular, we consider the Wasserstein distance between the two histogram distributions. The first histogram is constructed via probability scores predicted for each data instance, and the second histogram corresponds to the ideal scenario. Furthermore, given that the Wasserstein distance between both distributions does not consider the classifiers’ imperfection, we must weigh it with some metric that measures such imperfection. 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 propose two metrics, which we name Additive Probability Calibration Score (APCS - see Eq. 1) and Multiplicative Probability Calibration Score (MPCS - see Eq. 2). Both summarize the calibrated models’ performance, considering the classifier’s imperfection 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 density 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. 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. We detail the APCS equation in Eq. 1:
We note $APCS_{AUCROC} = |0.5 - AUCROC_{Classifier_{test}}|$. Here $AUCROC_{Classifier_{test}}$ corresponds to the classifiers’ AUC ROC measured on the test set, $W_1(h_i, h_{ref})$ is the 1-Wasserstein distance between the histogram $h_i$ and the reference histogram $h_{ref}$ and $n$ is the number of classes.

$$APCS = K_{APCS_{AUCROC}} + \sum_{i=1}^{n} \frac{1 - W_1(h_i, h_{ref})}{2 \cdot n} \tag{1}$$

On the other hand, MPCS corresponds 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 corresponds 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. We can express the MPCS equation as in Eq. 2.

$$MPCS = K_{MPCS_{AUCROC}} \sum_{i=1}^{n} \frac{1 - W_1(h_i, h_{ref})}{n} \tag{2}$$

Fig. 2: The figure illustrates three histograms that correspond to ideal cases described in this section: Perfect probability Calibration Model (PCM), Almost Perfect probability Calibration Model with perfect classification performance (APCM), and a probability Calibration Model with Perfect Classification performance, and Perfect Confidence (PCPCCM)
We consider three different cases of ideal target probability scores' histograms to measure the APCS and MPCS (see Fig. 2): (a) a Perfect probability Calibration Model (PCM), (b) an Almost Perfect probability Calibration Model with perfect classification performance (APCM), and (c) a probability Calibration Model with Perfect Classification performance and Perfect Confidence (PCPCCM). In PCM, the predicted probability value corresponds to the likelihood of occurrence of a given class, predicting values between zero and one. APCM has a similar shape, but positive occurrences for a given class occur for a probability score greater than $1/n$ where $n$ is the number of classes. Finally, PCPCCM produces prediction scores that can be translated to a single value (one) predicted for the correct class and zero otherwise. We consider it is not possible to obtain a perfect classification model with perfect probabilities calibration: given a probability score between zero and $1/n$, such cases must satisfy the condition that the likelihood of occurrence of a particular positive event for a given class is close to the probability score. Therefore, such events have a low probability score, resulting in misclassification and harming the model’s discriminative power. On the other hand, if a perfect classification performance is achieved, the probabilistic scores below $1/n$ cannot satisfy the probabilistic condition. Furthermore, we consider that a perfect probability calibration model cannot be achieved with a perfect and perfectly confident classifier, given that the first one emphasizes having a better spread of probability scores. In contrast, the second one attempts to assign the highest possible probability score to the predicted class. It must be noticed that the shape of such histograms is not influenced by class imbalance.

4 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. 3): double prints (the whole logo 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. However, given that such models are imperfect, the manual revision must 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, based on quality policies,
and the probability of obtaining a defective product given a particular prediction score. Furthermore, products sent to manual inspection can be prioritized using different criteria to enhance the existing defect detection machine learning model. In this research, we explore the aforementioned capabilities through multiple experiments, building supervised models, leveraging active learning, and comparing six different machine learning algorithms. We do so with a dataset of 3,518 labeled images, all of them corresponding to manufactured shavers.

5 Methodology

5.1 Active Learning strategies

We frame the automated defect detection as a supervised, multiclass classification problem. We used the ResNet-18 model [63] as a feature extractor, extracting 512 values long vectors for each image obtained from the Average Pooling layer. To avoid overfitting, we followed the procedure suggested by Hua et al. [64], and selected 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 kind of 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 [65].
To evaluate the models’ and active learning scenarios performance, we applied a stratified k-fold cross validation [66], considering $k=10$ based on recommendations by Kuhn et al. [67]. We used one fold for testing (test set), one for machine learning models’ probabilities calibration (calibration set), three folds to simulate a pool of unlabeled data for active learning (active learning set), and the rest to train the model (train set) (see Fig. 5). 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, we use two types of oracles: (a) machine oracles, which can be imperfect, and (b) human annotators, which we assume are ideal. We evaluated five machine learning algorithms: 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, we computed the AUC ROC metric. Given the multiclass setting, we used the ”one-vs-rest” heuristic, 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, we compared the AUC ROC values obtained by evaluating the model against the test fold for the first and last quartiles of instances queried in an active learning setting. We also assess the amount of manual work saved under each active learning setting and the precision of the soft-labeling approaches. Finally, we compare the probability calibration techniques to assess which was most advantageous for this particular use case.

Through different experiments, we aimed to simulate a visual inspection pipeline we expect to exist in a production setting (see Fig. 6). First, a stream of images is directed towards the machine learning model trained to identify possible defects. Then, based on the prediction score, a decision is made 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 (we consider it a soft label when not made by a human annotator). The label is then persisted, adding enlarging the existing dataset. The enlarged dataset can be used to retrain the model and replace the existing one after a successful deployment. We provide a detailed description of the experiments in Section 6.
5.2 Probability calibration strategies

To evaluate the probability calibration techniques, we follow a similar procedure described in the previous subsection but avoid the active learning step. Furthermore, we considered a different dataset split (see Fig. 7). After training the machine learning model and calibrating it with the calibration set, we issue a prediction for each instance of the unlabeled data set. We measure four performance metrics: AUC ROC, ECE, APCS, and MPCS. 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. We use it to compare the calibration quality for the multiple calibration techniques. Finally, APCS and MPCS estimate the calibration model’s quality without needing a ground truth. We contrast them against the above-mentioned metrics to measure the calibration methods’ performance without ground truth.
6 Experiments

6.1 Active Learning strategies

Fig. 8: 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.

For this research, we explored two active learning settings (pool-based and stream-based), using four distinct strategies to label the queried data instances in an active learning setting. We used two strategies 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 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, we analyzed the prediction for each data instance and derived it to the oracles for labeling if it was below a certain confidence threshold (p=0.95 or p=0.99).

We experimented with three oracle settings (see Fig. 8): (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 a set of three randomly picked images (one image 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 made based on the models’ confidence regarding the outcome and a probability threshold set based on manufacturing quality policies.

We set up eight scenarios (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. We calibrated the machine learning models using a sigmoid model based on Platt logistic model [68] (see Eq. 3).

**Equation 3** 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 parameters that are adjusted when fitting the regressor.

$$P(y_i = 1 \mid f_i) = \frac{1}{1 + exp(Af_i + B)}$$  (3)

| 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 |
6.2 Probability calibration strategies

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, we compare the models’ outcomes predictions 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), (b) temperature scaling (Temperature), (c) histogram calibration (with ground truth - we name it Histogram), (d) a histogram-based calibration approach (building the calibration map with the calibration set, but no ground truth - we name it AHPC (fixed)), (d) a histogram-based approach (building the calibration map with the calibration set and the rest of the predictions obtained from the unlabeled data on which predictions are issued, without ground truth (see Section 3) we name it AHPC (adaptive)).

In our experiments, we were interested in three aspects: (a) how calibration techniques compare against each other, (b) whether calibrating a model without a ground truth can provide comparable results to models calibrated with ground truth, and (c) if tracking the predictions of a given model can enhance the quality of the calibration model even when no ground truth is provided.

7 Results and Evaluation

7.1 Active Learning strategies

We analyzed the active learning strategies 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, we measured the AUC ROC over time (see Table 2). In particular, we contrast the models’ average performance 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, we found 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, we observed that 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 model’s performance over time. Nevertheless, further analysis
Table 2 Mean values for the mean AUC ROC computed across ten folds for five machine learning models. The results show how different active learning policies influence the models’ learning over time (Q1 vs. Q4). We consider two probability thresholds (0.95 and 0.99) as a cut-off for soft labeling. Best results are bolded, and second-best ones are displayed in italics.

| Setup       | Experiment | p=0.95       | p=0.99       |
|-------------|------------|--------------|--------------|
|             | Q1         | Q4           | Q1           | Q4           |
| Pool-based  | 1          | 0.8428       | 0.8612       | 0.8431       | 0.8623       |
|             | 2          | **0.8594**   | **0.8693**   | **0.8594**   | **0.8693**   |
|             | 3          | 0.8398       | 0.8396       | 0.8398       | 0.8398       |
|             | 4          | 0.8349       | 0.8348       | 0.8358       | 0.8358       |
|             | 5          | 0.8460       | 0.8559       | 0.8460       | 0.8559       |
| Streaming   | 6          | 0.8525       | 0.8647       | 0.8529       | 0.8647       |
|             | 7          | 0.8505       | 0.8608       | 0.8529       | 0.8647       |
|             | 8          | **0.8550**   | **0.8665**   | **0.8553**   | **0.8668**   |

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.

Table 3 Mean AUC ROC values computed across ten test folds for five machine learning models. 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)). Best results are bolded, and second-best results are displayed in italics.

| Model   | Q1             | Q4             | DS(p=0.95) |
|---------|----------------|----------------|------------|
| MLP     | **0.9309±0.0004** | **0.9448±0.0003** | Yes        |
| SVM     | **0.8788±0.0007** | **0.8767±0.0007** | 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        |

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 worst results were obtained by the CART model, which lagged a little more than 0.16 AUC ROC points from 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, whether some strategy can be devised (e.g., a machine oracle) to reduce the manual annotation effort. In this work, we focused on the second one. In Table 4, we present the results for a cut-off value of $p=0.95$. For $p=0.99$, we observed that no instances were retrieved and given to machine oracles; therefore, no analysis was performed on them.

When considering the cut-off value of 0.95, we noticed that 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, we 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.

Table 4 Proportion and quality of soft labeling through different settings, considering a predicted probability cut-off value of $p=0.95$. The task required annotating 2460 samples on average. $SL$ (%) 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.

| Experiment | $p=0.95$ | $SL$ (%) | $SL \ OK$ (%) | $ML \ SL \ OK$ (%) | $SSIM \ SL \ OK$ (%) |
|------------|----------|----------|---------------|---------------------|----------------------|
| 3          | 0.0077   | 0.9884   | 0.0075        | NA                  |
| 4          | 0.0033   | 0.9756   | 0.0090        | 0.0033              |
| 7          | 0.0143   | 0.9895   | 0.0480        | NA                  |
| 8          | 0.0144   | 0.9907   | 0.0484        | 0.0334              |
7.2 Probability calibration strategies

We present the results obtained when comparing the probability calibration techniques in two tables: Table 5 and Table 6. Table 5 describes the AUC ROC, ECE, and MPCS metric values obtained w.r.t. the three ideal histograms. Furthermore, we include the $MPCS - K_{MPCS_{AUCROC}}$ and $K_{MPCS_{AUCROC}}$ components to distinguish better how the models’ classification error and calibration imperfection influence the final MPCS scores. Similarly, Table 6 describes the AUC ROC, ECE, and APCS metric values obtained w.r.t. the three ideal histograms. Again, we include the $APCS - K_{APCS_{AUCROC}}$ and $K_{APCS_{AUCROC}}$ components, to distinguish better how the models’ classification error and calibration imperfection influence the final APCS scores.

In all but one case (for the CART model), the Histogram calibration (with ground truth) performed best when considering the ECE metric. The second best calibration technique was the Platt scaling, following the same pattern as the Histogram calibration technique. Furthermore, Platt scaling led to superior AUC ROC scores in three of five cases, with Histogram calibration (with ground truth) as second best in three of five cases. Using the AHPC technique did not strongly impact the AUC ROC metric, providing enhancements or a slight detriment to the final models’ discriminative power no larger than 0.04 AUC ROC points. When considering the calibration quality, we found that the AHPC technique had the worst ECE scores in all cases but one and was even surpassed by the performance obtained by models’ scores alone, without calibration. The best performance was obtained by the Histogram calibration (with ground truth) and the second best by the Platt scaling technique.

Analyzing the values obtained for the MPCS metrics, we found that for the PCPCCM setting, the best and second-best results corresponded to the models with the highest AUC ROC values and usually the best or second-best ECE score. While we expected a strong influence of the ROC AUC score on the MPCS metric for the PCPCCM case, we were surprised to observe a correlation with the ECE score, given the difference between the target shape of the ideal case pursued by the ECE metric and the PCPCCM histogram. Results w.r.t. ideal histograms PCM and APCM differed w.r.t. the PCPCCM case. While MPCS for PCM and APCM identified the same cases as best-performing models, minor discrepancies were observed for the second-best ones. We consider the metrics reflected our expectations regarding the MPCS behavior for both cases, given the similarity between both ideal histograms. The discrepancies for the second-best models only reinforced the usefulness of both metrics to assess the behavior of the calibration techniques w.r.t. the desired outcome.

When considering the results obtained for the APCS metrics, we did not find a clear winner among the calibration techniques.

Analyzing the values obtained for the MPCS metrics, we found that for the PCPCCM setting, the best and second-best results corresponded to the models with the highest AUC ROC values and usually the best or second-best ECE score. While we expected a strong influence of the ROC AUC score on the MPCS metric for the PCPCCM case, we were surprised to observe a correlation with the ECE score, given the difference between the target shape of the ideal case pursued by the ECE metric and the PCPCCM histogram. Results w.r.t. ideal histograms PCM and APCM differed w.r.t. the PCPCCM case. While MPCS for PCM and APCM identified the same cases as best-performing models, minor discrepancies were observed for the second-best ones. We consider the metrics reflected our expectations regarding the MPCS behavior for both cases, given the similarity between both ideal histograms. The discrepancies for the second-best models only reinforced the usefulness of both metrics to assess the behavior of the calibration techniques w.r.t. the desired outcome.

When considering the results obtained for the APCS metrics, we did not find a clear winner among the calibration techniques.
the MPCS metric for the PCPCCM case, we were surprised to observe a correlation with the ECE score, given the difference between the target shape of the ideal case pursued by the ECE metric and the PCPCCM histogram. Results w.r.t. ideal histograms PCM and APCM differed when compared against the PCPCCM case. Depending on the machine learning model, we observed that Platt scaling, Temperature calibration, or AHPC were considered best. While MPCS for PCM and APCM identified the same best-performing models, minor discrepancies were observed for the second-best ones. We consider the measured values reflected our intuition regarding the MPCS behavior for both cases, given the similarity between both ideal histograms. The discrepancies for the second-best models only reinforced the usefulness of both metrics in assessing the behavior of the calibration techniques against the desired outcome.

To understand how the APCS and MPCS reflect the information present at the ECE score (computed with ground truth), we computed the Pearson correlation between ECE and the APCS or MPCS scores. For the APCS score, we found that the strongest correlation is displayed for the PCM setting (Pearson correlation of 0.08), while it weakened by at least 0.03 points for the remaining two settings. Furthermore, the correlation between ECE and the component described by the Wasserstein distance ($\text{APCS} - K_{\text{APCS} \text{AUCROC}}$) was 0.39 for the PCPCCM setting, followed by the rest of the settings by at least 0.03 points. Considering the case of the MLP model alone, which achieved the highest AUC ROC results, we validated that the correlations were more pronounced. Such strong correlations signal that information regarding the ground truth measured by the ECE metric was also captured by the APCS metric while not requiring such a ground truth and relying on the model’s predictions alone. For the particular case of the MLP model, we measured a Pearson correlation between 0.58 and 0.74. Again, the highest correlation was measured for the PCM case. Furthermore, when considering only the component described by the Wasserstein distance ($\text{APCS} - K_{\text{APCS} \text{AUCROC}}$), PCM and APCM displayed a correlation higher than 0.99, while for PCPCCM, we measured a correlation value of -0.59.

For the MPCS score, we found that the strongest correlation with the ECE metric was measured for the PCM setting (Pearson correlation of 0.13). The second-highest correlation was measured for the APCM setting (0.10), while the PCPCCM almost did not correlate with the ECE metric. The highest correlation between ECE and the component described by the Wasserstein distance ($\text{APCS} - K_{\text{MPCS} \text{AUCROC}}$) was 0.35 (PCM setting), followed by the rest of the settings by at least 0.06 points. Considering the case of the MLP model alone (highest AUC ROC among the models), the correlation between ECE and MPCS was 0.81 for PCM, 0.74 for APCM, and -0.35 for PCPCCM. Furthermore, when considering only the component described by the Wasserstein distance ($\text{APCS} - K_{\text{APCS} \text{AUCROC}}$), PCM and APCM displayed a correlation higher than 0.98, while for PCPCCM, we measured a correlation value of -0.53.
When considering the AHPC (adaptive) setting, we observed that in two cases the quality of calibration improved over time (achieved lower ECE after adjusting the histogram values with the "unlabeled" data set), in two cases it worsened, and in one case it remained the same. Therefore we conclude that further experimentation is required to understand if such a method can get competitive results over time against the rest of the methods and whether the adaptive setting can be beneficial in the face of concept drift.

While calibrating the model without a ground truth (AHPC in fixed or adaptive mode) did not provide comparable results to existing calibration techniques, we consider further work is required to understand (a) whether the technique is better suited to certain machine learning models, (b) if a greater amount of unlabeled data can enhance the final performance of such a technique and make its’ performance comparable to existing methods that require a ground truth, and (c) whether the AHPC method can provide some level of probability calibration and resilience when concept drift takes place.

8 Conclusions and Future Work

In this work, we 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 $p=0.95$. Furthermore, we compared multiple probability calibration techniques, proposed a probability calibration technique that does not require a ground truth, and two metrics to measure the quality of the model, summarizing its capacity to discriminate between classes and to produce adequate probability scores. Through the experiments, we 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. Furthermore, the APCS and MPCS metrics capture such information without needing ground truth data.

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) and enhancing the calibration techniques to display the desired behavior for high-confidence thresholds. Regarding probabilities calibration, we will conduct further research regarding the AHPC technique to understand how it can be improved and if using a larger amount of predicted scores can enhance its performance to achieve competitive results against other probability calibration techniques. Finally, we will explore how robust the proposed technique is against concept drift and whether it provides advantages against other calibration techniques in such a setting.

Acknowledgments. This work was supported by the Slovenian Research Agency and the European Union’s Horizon 2020 program project STAR under grant agreement number H2020-956573.
Table 5  The results were obtained for different models and probability calibration techniques. Best results (per model) are bolded, and second-best are shown in italics. We measure the models’ discriminative power (AUC ROC), quality of calibration considering a ground truth (ECE), and quality of calibration without considering a ground truth for three ideal target states (MPCS for histograms PCM, APCM, and PCPCCM). We decompose the MPCS metric reporting the $K_{\text{MPCS}_{\text{AUCROC}}}$ and $PCS - K_{\text{MPCS}_{\text{AUCROC}}}$, to inform better how the calibration and models’ imperfection contribute to the final score.

| Model | Calibration | AUC ROC | ECE | PCM | PCPCCM | PCM | PCPCCM | APCM | PCM | PCPCCM |
|-------|-------------|---------|-----|-----|--------|-----|--------|-------|-----|--------|
| CART  | No calibration | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | Histogram | 0.7363  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | Temperature | 0.7471  | 0.5215 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 |
|       | AIPC (fixed) | 0.7410  | 0.5518 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 |
|       | AIPC (adaptive - start) | 0.7350  | 0.5518 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 |
|       | AIPC (adaptive - end) | 0.7310  | 0.5518 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 |
|       | No calibration | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | Histogram | 0.7363  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | Temperature | 0.7471  | 0.5215 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 |
|       | AIPC (fixed) | 0.7410  | 0.5518 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 |
|       | AIPC (adaptive - start) | 0.7350  | 0.5518 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 |
|       | AIPC (adaptive - end) | 0.7310  | 0.5518 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 | 0.3402 |
|       | No calibration | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | Histogram | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | Temperature | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | AIPC (fixed) | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | AIPC (adaptive - start) | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | AIPC (adaptive - end) | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | No calibration | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | Histogram | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | Temperature | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | AIPC (fixed) | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | AIPC (adaptive - start) | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |
|       | AIPC (adaptive - end) | 0.7460  | 0.6060 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 | 0.3382 |

This document is the property of the STAR consortium and shall not be distributed or reproduced without the formal approval of the STAR Management Committee. The content of this report reflects only the authors’ view. The European Commission is not responsible for any use that may be made of the information it contains.

**Declarations**

**Conflict of interest**

The authors have no competing interests to declare that are relevant to the content of this article.
Table 6 The results were obtained for different models and probability calibration techniques. Best results (per model) are bolded, and second-best are shown in italics. We measure the models’ discriminative power (AUC ROC), quality of calibration considering a ground truth (ECE), and quality of calibration without considering a ground truth for three ideal target states (APCS for histograms PCM, APCM, and PCPCCM). We decompose the APCS metric reporting the $K_{\text{APCS}_{\text{AUCROC}}}$ and $PCS - K_{\text{APCS}_{\text{AUCROC}}}$, to inform better how the calibration and models’ imperfection contribute to the final score.

| Model | Calibration | AUC ROC | ECE | PCM | APCM | $K_{\text{APCS}_{\text{AUCROC}}}$ | $PCS - K_{\text{APCS}_{\text{AUCROC}}}$ |
|-------|-------------|---------|-----|-----|------|----------------------------------|----------------------------------------|
| CART  | No calibration | 0.7562 | 0.0199 | 0.1692 | 0.2760 | 0.8772 |
|       | Histogram    | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Platt        | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Temperature  | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (fixed) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - start) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - end) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
| KNN   | No calibration | 0.7562 | 0.0199 | 0.1692 | 0.2760 | 0.8772 |
|       | Histogram    | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Platt        | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Temperature  | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (fixed) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - start) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - end) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
| NB    | No calibration | 0.7562 | 0.0199 | 0.1692 | 0.2760 | 0.8772 |
|       | Histogram    | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Platt        | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Temperature  | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (fixed) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - start) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - end) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
| MLP   | No calibration | 0.7562 | 0.0199 | 0.1692 | 0.2760 | 0.8772 |
|       | Histogram    | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Platt        | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Temperature  | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (fixed) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - start) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - end) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
| SVM   | No calibration | 0.7562 | 0.0199 | 0.1692 | 0.2760 | 0.8772 |
|       | Histogram    | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Platt        | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | Temperature  | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (fixed) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - start) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |
|       | ARPC (adaptive - end) | 0.7562 | 0.0222 | 0.1692 | 0.2760 | 0.8772 |

Availability of data and materials
The datasets analysed during the current study are not publicly available due confidentiality reasons.

References
[1] Yang, J., Li, S., Wang, Z., Dong, H., Wang, J., Tang, S.: Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges. Materials 13(24), 5755 (2020)
[2] Kurniati, N., Yeh, R.-H., Lin, J.-J.: Quality inspection and maintenance: the framework of interaction. Procedia manufacturing 4, 244–251 (2015)
[3] Wuest, T., Irgens, C., Thoben, K.-D.: An approach to monitoring quality in manufacturing using supervised machine learning on product state data. Journal of Intelligent Manufacturing 25(5), 1167–1180 (2014)

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

[5] Zheng, Z., Zhang, S., Yu, B., Li, Q., Zhang, Y.: Defect inspection in tire radiographic image using concise semantic segmentation. IEEE Access 8, 112674–112687 (2020)

[6] Vergara-Villegas, O.O., Cruz-Sánchez, V.G., Jesús Ochoa-Domínguez, H.d., Jesús Nandayapa-Alfaro, M.d., Flores-Abad, Á.: Automatic product quality inspection using computer vision systems. In: Lean Manufacturing in the Developing World, pp. 135–156. Springer, ??? (2014)

[7] Selvi, S.S.T., Nasira, G.: An effective automatic fabric defect detection system using digital image processing. J. Environ. Nanotechnol 6(1), 79–85 (2017)

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

[9] Cullinane, S.-J., Bosak, J., Flood, P.C., Demerouti, E.: Job design under lean manufacturing and its impact on employee outcomes. Organizational Psychology Review 3(1), 41–61 (2013)

[10] Kujawińska, A., Vogt, K., Hamrol, A.: 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, ??? (2016)

[11] Rai, R., Tiwari, M.K., Ivanov, D., Dolgui, A.: Machine learning in manufacturing and industry 4.0 applications. Taylor & Francis (2021)

[12] Zheng, T., Ardolino, M., Bacchetti, A., Perona, M.: The applications of industry 4.0 technologies in manufacturing context: a systematic literature review. International Journal of Production Research 59(6), 1922–1954 (2021)

[13] Rožanec, J.M., Novalija, I., Zajec, P., Kenda, K., Tavakoli, H., Suh, S., Veliou, E., Papamartzivanos, D., Giannetsos, T., Menesidou, S.A., et al.: Human-centric artificial intelligence architecture for industry 5.0 applications. arXiv preprint arXiv:2203.10794 (2022)
Active Learning and Approximate Model Calibration in Manufacturing

[14] Carvajal Soto, J., Tavakolizadeh, F., Gyulai, D.: 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 (2019)

[15] Chouchene, A., Carvalho, A., Lima, T.M., Charrua-Santos, F., Osório, G.J., Barhoumi, W.: 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 (2020). IEEE

[16] Duan, G., Wang, H., Liu, Z., Chen, Y.-W.: 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 (2012)

[17] Jiang, J., Wong, W.: Fundamentals of common computer vision techniques for textile quality control. In: Applications of Computer Vision in Fashion and Textiles, pp. 3–15. Elsevier, ??? (2018)

[18] Villalba-Diez, J., Schmidt, D., Gevers, R., Ordieres-Meré, J., Buchwitz, M., Wellbrock, W.: Deep learning for industrial computer vision quality control in the printing industry 4.0. Sensors 19(18), 3987 (2019)

[19] Beltrán-González, C., Bustreo, M., Del Bue, A.: External and internal quality inspection of aerospace components. In: 2020 IEEE 7th International Workshop on Metrology for AeroSpace (MetroAeroSpace), pp. 351–355 (2020). IEEE

[20] 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.: Artificial intelligence/machine learning in manufacturing and inspection: A ge perspective. MRS Bulletin 44(7), 545–558 (2019)

[21] Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P.: Classifier calibration: How to assess and improve predicted class probabilities: a survey. arXiv e-prints, 2112 (2021)

[22] Rožanec, J.M., Trajkova, E., Dam, P., Fortuna, B., Mladenić, D.: Streaming machine learning and online active learning for automated visual inspection. IFAC-PapersOnLine 55(2), 277–282 (2022)

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

[24] Schmitt, J., Bönig, J., Borggräfe, T., Beitingler, G., Deuse, J.: Predictive
model-based quality inspection using machine learning and edge cloud computing. Advanced engineering informatics 45, 101101 (2020)

[25] Weiss, S.M., Dhurandhar, A., Baseman, R.J., White, B.F., Logan, R., Winslow, J.K., Poindexter, D.: Continuous prediction of manufacturing performance throughout the production lifecycle. Journal of Intelligent Manufacturing 27(4), 751–763 (2016)

[26] Park, J.-K., Kwon, B.-K., Park, J.-H., Kang, D.-J.: Machine learning-based imaging system for surface defect inspection. International Journal of Precision Engineering and Manufacturing-Green Technology 3(3), 303–310 (2016)

[27] Ren, R., Hung, T., Tan, K.C.: A generic deep-learning-based approach for automated surface inspection. IEEE transactions on cybernetics 48(3), 929–940 (2017)

[28] Jian, C., Gao, J., Ao, Y.: Automatic surface defect detection for mobile phone screen glass based on machine vision. Applied Soft Computing 52, 348–358 (2017)

[29] Tsai, D.-M., Lai, S.-C.: Defect detection in periodically patterned surfaces using independent component analysis. Pattern Recognition 41(9), 2812–2832 (2008)

[30] Kang, G.-W., Liu, H.-B.: 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 (2005). IEEE

[31] Yun, J.P., Choi, D.-c., Jeon, Y.-j., Park, C., Kim, S.W.: Defect inspection system for steel wire rods produced by hot rolling process. The International Journal of Advanced Manufacturing Technology 70(9), 1625–1634 (2014)

[32] Valavanis, I., Kosmopoulos, D.: Multiclass defect detection and classification in weld radiographic images using geometric and texture features. Expert Systems with Applications 37(12), 7606–7614 (2010)

[33] Cohen, N., Hoshen, Y.: Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357 (2020)

[34] Li, C.-L., Sohn, K., Yoon, J., Pfister, T.: 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 (2021)
Active Learning and Approximate Model Calibration in Manufacturing

[35] Jezek, S., Jonak, M., Burget, R., Dvorak, P., Skotak, M.: 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 (2021). IEEE

[36] Aiger, D., Talbot, H.: The phase only transform for unsupervised surface defect detection. In: Emerging Topics In Computer Vision And Its Applications, pp. 215–232. World Scientific, ??? (2012)

[37] Mujeeb, A., Dai, W., Erdt, M., Sourin, A.: Unsupervised surface defect detection using deep autoencoders and data augmentation. In: 2018 International Conference on Cyberworlds (CW), pp. 391–398 (2018). IEEE

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

[39] Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine Learning Proceedings 1994, pp. 148–156. Elsevier, ??? (1994)

[40] Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine learning 15(2), 201–221 (1994)

[41] Settles, B.: Active learning literature survey (2009)

[42] Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., Wang, X.: A survey of deep active learning. arXiv preprint arXiv:2009.00236 (2020)

[43] Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: 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 (2018)

[44] Cordier, A., Das, D., Gutierrez, P.: Active learning using weakly supervised signals for quality inspection. arXiv preprint arXiv:2104.02973 (2021)

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

[46] Dai, W., Mujeeb, A., Erdt, M., Sourin, A.: Towards automatic optical inspection of soldering defects. In: 2018 International Conference on
Cyberworlds (CW), pp. 375–382 (2018). IEEE

[47] Zajec, P., Rožanec, J.M., Novalija, I., Fortuna, B., Mladenić, D., Kenda, K.: Towards active learning based smart assistant for manufacturing. In: IFIP International Conference on Advances in Production Management Systems, pp. 295–302 (2021). Springer

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

[49] Cheeseman, P.C.: In defense of probability. In: IJCAI, vol. 85, pp. 1002–1009 (1985)

[50] Cohen, I., Goldszmidt, M.: Properties and benefits of calibrated classifiers. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 125–136 (2004). Springer

[51] Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P., et al.: Classifier calibration: How to assess and improve predicted class probabilities: a survey. arXiv preprint arXiv:2112.10327 (2021)

[52] Kumar, A., Liang, P.S., Ma, T.: Verified uncertainty calibration. Advances in Neural Information Processing Systems 32 (2019)

[53] Platt, J.C.: 5 probabilities for sv machines. Advances in Large Margin Classifiers, 61 (2000)

[54] Leathart, T., Frank, E., Holmes, G., Pfahringer, B.: Probability calibration trees. In: Asian Conference on Machine Learning, pp. 145–160 (2017). PMLR

[55] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017). PMLR

[56] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems 32 (2019)

[57] Bröcker, J., Smith, L.A.: Increasing the reliability of reliability diagrams. Weather and forecasting 22(3), 651–661 (2007)

[58] Gupta, C., Ramdas, A.: Distribution-free calibration guarantees for histogram binning without sample splitting. In: International Conference on Machine Learning, pp. 3942–3952 (2021). PMLR

[59] Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring
calibration in deep learning. In: CVPR Workshops, vol. 2 (2019)

[60] Dunning, T.: The t-digest: Efficient estimates of distributions. Software Impacts 7, 100049 (2021)

[61] Villani, C.: Optimal Transport: Old and New vol. 338. Springer, ??? (2009)

[62] Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning 11(5-6), 355–607 (2019)

[63] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[64] Hua, J., Xiong, Z., Lowey, J., Suh, E., Dougherty, E.R.: Optimal number of features as a function of sample size for various classification rules. Bioinformatics 21(8), 1509–1515 (2005)

[65] Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural computing and applications 24(1), 175–186 (2014)

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

[67] Kuhn, M., Johnson, K., et al.: Applied Predictive Modeling vol. 26. Springer, ??? (2013)

[68] Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10(3), 61–74 (1999)