Finding optimal decision boundaries for human intervention in one-class machine-learning models for industrial inspection

Tim Zander[*] Ziyan Pan Pascal Birnstill
Jürgen Beyerer[†]

Abstract

Anomaly detection with machine learning in industrial inspection systems for manufactured products relies on labelled data. This raises the question of how the labelling by humans should be conducted. Moreover, such a system will most likely always be imperfect and potentially need a human fall-back mechanism for ambiguous cases. We consider the case where we want to optimise the cost of the combined inspection process done by humans together with a pre-trained algorithm. This gives improved combined performance and increases the knowledge of the performance of the pre-trained model. We focus on so-called one-class classification problems which produce a continuous outlier score. After establishing some initial setup mechanisms ranging from using prior knowledge to calibrated models, we then define some cost model for machine inspection with a possible second inspection of the sample done by a human. Further, we discuss in this cost model how to select two optimal boundaries of the outlier score, where in between these two boundaries human inspection takes place. Finally, we frame this established knowledge into an applicable algorithm and conduct some experiments for the validity of the model.

Mathematical methods and models, artificial intelligence and machine learning, quality control, active learning, label effort, one-class classification

1 Introduction

The detection of non-common patterns in a batch of samples is a strong point of human visual cognition. Still, there are many known limitations to human visual inspection as well as cost issues and labour shortages in

[*] Institut für Anthropomatik und Robotik, Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie (KIT), Karlsruhe, Germany, e-mail: tim.zander@kit.edu

[†] Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (IOSB), Karlsruhe, Germany

real-world production systems. The training of machine learning models for anomaly detection of industrial inspection problems is often done as a one-class classification problem where only good samples are presented to the algorithm. The background for this is that it is generally easy to acquire good samples, but difficult and expensive to find anomalous samples. A data set for benchmarking this type of algorithm is the MVTec data set[1][2]. The best performing model[1] on this data set to-date is *"Patchcore"*[20]. Although we think of models designed for the MVTec data set like *"Patchcore"* as the main application, our method of finding two boundaries for the outlier score, where in between an additional round of human inspection will take place, will work for any model for one-class classification problems[17] with a continuous score.

More precisely, in this paper, we formulate the problem of optimal usage of human inspection for edge cases after acquiring initial data for training. For this, we assume that there are certain costs for inspection and costs for falsely classified samples. We are not aware that such a human-in-the-loop machine learning consideration exists in the literature, although more generic considerations about iterative machine teaching and active learning can be found in [12]. A similar process by giving the human some sort of optimal presentation of data for labelling was done in [3]. However, this method does not apply to the one-class outlier classification problems on images which we do consider here. In [21] it is shown, that for one-class classification models, one can train an additional model on the bad samples and use a combined score on the good and bad sample models to find the most promising new samples for labelling. The authors show that by using one of their active learning methods one can achieve faster convergence and better overall performance of the model. We refer to Munro's book [13] for a general overview of human-in-the-loop machine learning and active learning in particular.

In essence, the method we develop will separate samples after passing through an outlier detection algorithm into three categories depending on their outlier score. The ones which will be immediately accepted, the ones which will be immediately rejected and the ones which will be given to a human for manual inspection. Such a distinction is especially relevant for inspection tasks where computer vision outlier detection does not perform sufficiently well. Here, a combination of human and computer vision algorithms aims to provide better results for industrial inspection tasks in terms of cost. Such hybrid inspection approaches being viable has been known for a long time [11, 10], however, to our knowledge, no concrete decision boundaries for human inspection for the algorithms were established.

As an example inspection task, we take a production line where it is not feasible to inspect every piece by a human. We take an artificial example such as the Utah teapot (see Figure 1) and will show the edge cases to the user, where they decide whether an anomaly is present or not. We will also test various future hybrid approaches involving visual inspection in an experimental set-up we call **HALODOME** (*HumAn in the Loop Outlier DetectiOn MachinE*). The idea is to simulate the machine

[1] `https://paperswithcode.com/sota/anomaly-detection-on-mvtec-ad`

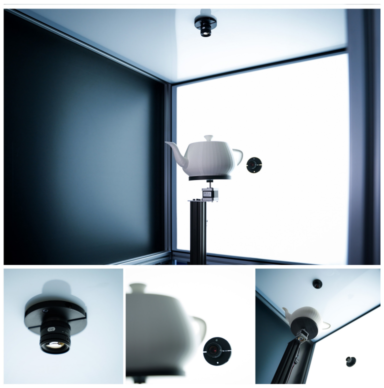


Figure 1: These are pictures of our HALODOME-setup which ought to simulate inline inspection of manufactured pieces.

teaching and later the edge case inspection done by a human, among other tasks such as 3D error marking and visualisation (see Figure 2).

2 Related Work

As already mentioned the best-performing algorithm for the MVtec data set is the so-called *"Patchcore"* model. For a given picture sample a *"Patchcore"*-model after training produces an outlier score together with a heat map on the likelihood of being an anomalous area (see Figure 3 for a visualisation of such a heat map). This is done by performing outlier detection on the deep features of a pre-trained neural network of the images. The cutoff values for an anomaly in the outlier score of *"Patchcore"* are optimised in the paper by finding the cut-off value with the highest F1-score. This already assumes that there are known outliers which are potentially very costly to acquire. We will later compare the cost performance of naive F1-optimisation of this algorithm with our approach.

Another important concept which we will discuss and use is that of probabilistic classifiers. Probabilistic classifiers are classifiers that output a probability distribution on the target classes instead of just a score. A calibrated classifier will give out a probability p which will represent the probability of being in a particular class. Machine-learning models M such as neural networks often already provide a likelihood with their output. The question which arises with such models is whether these likelihood outputs already resemble probabilistic classifiers and how good these probabilistic classifiers are. Model calibration is a technique which achieves that a classifier will have a probabilistic output[25] [6]. A measure for the quality of the likelihood as a probabilistic classifier is that of calibration.

Definition 1 (Rough definition)**.** *A machine learning model's prediction*

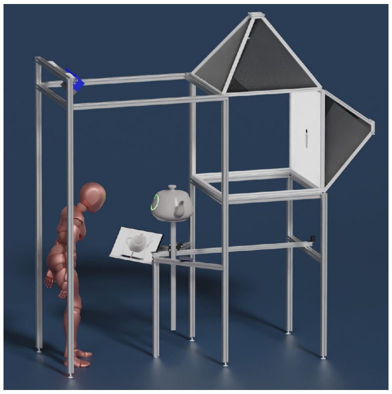
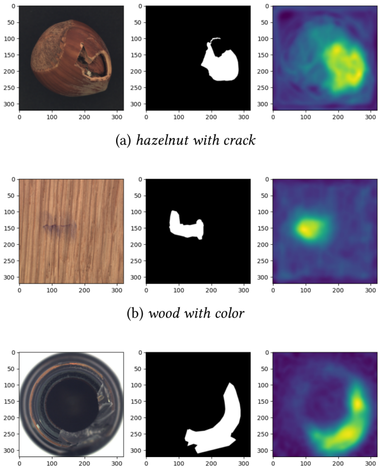


Figure 2: This shows a possible experimental setup in HALODOME. The data acquisition of the object, here represented as the Utah teapot, will take place in the box on the right. The object will be moved into the box by a linear robot. If an outlier is found in the sample it can be presented to the worker on the left either on display or via a projector mounted on the upper left directly on the object. Studies to simulate edge case inspection and helpfulness of error augmentation are planned in such a setup.



(a) *hazelnut with crack*

(b) *wood with color*

(c) *bottle with breakage*

Figure 3: A sample visualisation of trained *"Patchcore"*-models output. The left images are some test samples. The right images represent the pixel-wise outlier score, where the values are encoded in the visible spectrum between blue and yellow. The lowest values are encoded in dark blue and the highest in yellow. The middle image is a binarisation of the left where the cut-off parameter is chosen to be the F1-optimal boundary.

is calibrated when the likelihood output resembles the probability of its prediction being correct.

Definition 2 (Perfect calibration)**.** *Let $X \in \mathcal{X}$ (samples) and $Y \in \mathcal{Y}$ (classes) be random variables with joint distribution $\pi(X, Y) = \pi(Y|X) \cdot \pi(X)$. Let h be some machine learning model such that $h(X) = (\hat{Y}, \hat{P})$ with $\hat{Y}$ the class prediction and $\hat{P}$ the predicted probability of correctness. Now h is called* calibrated *if*

$$P(\hat{Y} = Y|\hat{P} = p) = p \text{ for all } p \in [0, 1]. \tag{1}$$

Of course, in reality, we have no access to the joint distribution and instead, check for p being in one of M-many bins B_m which contains all elements with predictions in the interval $(\frac{m-1}{M}, \frac{m}{M}]$. We then compare whether the average of the predicted p in each bin (this is called the confidence of B_m - $\text{conf}(B_m)$) equals the percentage of labelled data being correct in the bin (this is called the accuracy of B_m - $\text{acc}(B_m)$). As a measurement of the calibration in a formula, we have the *Expected Calibration Error*[15]

$$\sum_{m=1}^{M} \frac{|B_m|}{n} |\text{conf}(B_m) - \text{acc}(B_m)| \tag{2}$$

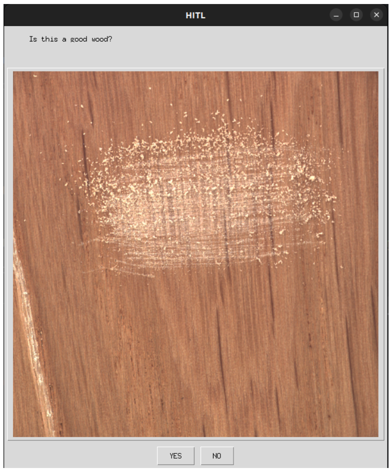
where n is the number of samples. Hence as in many applications, it is important to have an idea of the uncertainty of the model for which the expected calibration error provides a measure. In situations which are cost-critical, we will show that we can exploit having an uncertainty estimate of the classifier for a given sample to make better decisions. In fact, we will implicitly find some external calibration for the model through the definition of some statistical model.

3 Model

First, in this section, we will describe the necessary pre-conditions and cost assumptions. Further, we describe how, after initial training of our one-class classifier, we can establish our first optimal boundaries. We do describe multiple alternatives here. Then we pass on to acquiring more knowledge about the outliers which we will encounter and their outlier scores. This will then later be used to establish optimal decisions for the cut-off parameters of human inspection in the sense of our pre-made cost assumptions.

3.1 Pre-conditions

First, we introduce a few more preliminary and formal assumptions and notations. We assume that there exists a set of images or more general data I which each has a hidden label $\{0, 1\}$ where images with label 0 are good samples and images with label 1 are anomalous samples. We will observe these samples in some processes such as an industrial inspection task one after another. For our cost considerations we assume that the process of labelling a sample by a human has a cost c_l associated with



Is this a good wood?

Figure 4: A simple binary labelling UI. The use case is a wood image from MVTec AD[2].

it. Further, we assume that human labelling perfectly assigns the correct label to the data. Whether this is justified in reality for the MVTec data set could be tested with a simple user study using a very simple program as depicted in Figure 4. With N initially labelled data points, we train and test a model M which will then produce an outlier-score $M(i) \in \mathbb{R}$ for every (new) image i we observe. We set a lower and upper decision boundary for manual inspection b_l and b_u such that any image i with outlier score $M(i)$, where $b_l < M(i) < b_u$ holds, will be inspected by a human.

3.2 Inspection and fault cost and anomalous data

For our cost considerations, we assume that there is a known (possibly non-linear[2]) cost-function C_f such that the absolute cost of missed outliers can be calculated as $C_f(\mathbf{FOR}) \cdot K$ where **FOR** is the false omission rate, i.e. the percentage of anomalies in the accepted samples, and K is the absolute number of accepted samples. The cost of false positive samples is associated with a cost per sample of c_r. This could be for example lost revenue and disposal costs of an unnecessarily discarded sample in a good state.

3.3 Optimal inspection without human intervention

Even in case no human inspection is planned we can still apply the rest of the analysis done in this paper. This will then be an easier sub-problem

[2] One reason for non-linearity could be reputation costs, i.e., due to network effects reputation falls non-linearly with increasing fault-rate.

contained in our analysis, i.e., the problem of finding a single optimal cut-off value for outliers. This is always a potential solution to a general problem we will formulate later, where inspection by humans for some subset is considered. In particular, this means that the inspection interval has width 0, i.e. $b_l = b_u$.

3.4 Initial cut-off boundaries

We assume now that the initial sampling and labelling of data D and the training of a model M are conducted. We update our initial belief p_o of the outlier percentage by taking the percentage of outliers in the sampled D into account. We are now interested in finding optimal cut-off parameters b_l, b_u in this stage. We discuss multiple alternatives now.

3.4.1 A priori anomaly distribution

In the first case, we assume that the distribution of the outlier score of samples with label 0 and also of the samples with label 1 is both Gaussian[3]. For the good samples, we can directly estimate this distribution after observing our initial training data. We get some distribution g_g with mean μ_g and variance σ_g. For the bad samples, we also get some Gaussian distribution g_b (see Figure 5 for examples of such distributions). In the case where there are no bad samples available, we take some initial belief about the distribution, which we could take from former observations such as the MVTec data set or a similar product line as our distribution. We can find the optimal parameters b_l, b_u in terms of costs. In order to find these parameters one would minimise Equation 4 of Section 3.5.

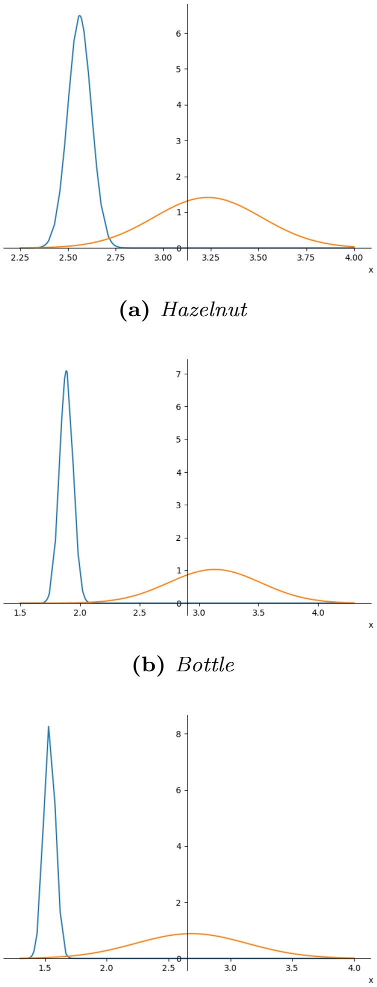
3.4.2 Optimal cut-off sigma

Another approach would be to omit to define an a priori distribution of g_b and instead, take a cut-off parameter x such that any sample with an outlier score higher than $\mu_g + x \cdot \sigma_g$ is considered anomalous. The choice of the parameter x can be done as follows. We assume that we cannot inspect every piece which we observe but only some percentage p_i of it. Hence we have to find x in such a way that the expected amount of samples classified as anomalous is at most the amount that can be handled. Hence we have to pick x such that

$$p_i \geq (1 - p_o) \int_{\mu_g + x \cdot \sigma_g}^{\infty} g_g(y)\, dy + p_o \tag{3}$$

holds. Note that we omitted the expected false negative classified samples in our considerations, but we assume that this amount is negligibly small. In case there is no sample to classify at the moment we might pick a random sample. In case we acquire enough bad samples we can infer the distribution g_b or update our initial belief about it. More details on the belief update of Gaussian distributions can be found in [14].

[3] A non-Gaussian distribution could also easily be considered here.



(a) *Hazelnut*

(b) *Bottle*

(c) *Leather*

Figure 5: These are the Gaussian distributions of anomaly scores for different items from the MVTec data set. The blue graphs represent the good sample distributions and the orange graphs represent the bad sample distributions. The model where the anomaly score stems from was Patchcore[20] and it was trained with a training sample split of the MVTec data set. Then the anomaly score output of the trained model on the good and bad samples of the test data set split was used to find the shown Gaussian distributions. On these data sets the established model has an AUC-score of 0.9996 for *Hazelnut*, 1.0 for *Bottle* and 1.0 for *Leather* on the test data set samples.

3.4.3 Calibrated output

In some cases, the model comes with a calibrated probabilistic output. As discussed before this roughly means that the output value of the model $M(*)$ is a probabilistic classifier, i.e. a sample x has probability $M(x)$ of being an outlier. With such a calibrated model we can directly use the model output as our probability for the conditional distribution $\mathcal{P}(Y|X)$. We will later see that the cost equation 4 defined in the Section 3.5 will be simplified under certain assumptions as we can observe the data on a point-to-point basis only considering the probabilistic output $\mathcal{P}(Y|X)$. As in our case, for now, we have in most cases a one-class classification problem so we need calibrated one-class classification models. For support vector machines there is existing work which conducts calibration of such classifiers in one-class classification problems[23, 9].

3.4.4 Conclusion for the setting of initial cut-off parameters

Now we have found prior parameters b_l, b_u or just $b_l(= \mu_g + x \cdot \sigma_g)$. With these, we can set up our initial human-in-the-loop process. After some time we will enrich our data set of labelled pieces and therefore can update our belief about the Gaussian curves g_g, g_b as described in [14] or interfere the distributions g_g, g_b directly from all the gathered data. There is some caveat with the selection of the samples: Because of our parameters, the selection of the samples is biased. This either needs to be corrected through enough random samples or giving the unlabelled data some pseudo label with a continuous value greater than 0 and smaller than 1. Additionally, we could use the gathered data to further improve the model M or respectively re-train a new M with the new data and old data depending on the algorithm in use. In any case, we now fix some model M, some p_0 and the Gaussian distributions g_g, g_b associated with it as well as the gathered data. In case we observed and classified a new sample we could continue to do a belief update of our estimated values p_o, g_g and g_b and retrain our model M to keep improving it. But we omit such considerations in the rest of the paper.

3.5 Cost-calculation

We calculate the cost associated for some fixed b_l and b_u for the next samples. We expect to see p_o-percent outliers which we have updated from the observations D. Additionally, we can calculate the expected percentage that the next sample will be true positive: $\mathbf{TP}(b_l) = p_o \int_{b_l}^{\infty} g_b(x)\,dx$, true negative: $\mathbf{TN}(b_u) = (1 - p_o) \int_{-\infty}^{b_u} g_g(x)\,dx$, false negative: $\mathbf{FN}(b_l) = p_o \int_{-\infty}^{b_l} g_b(x)\,dx$ and false positive: $\mathbf{FP}(b_u) = (1 - p_o) \int_{b_u}^{\infty} g_g(x)\,dx$. From this we can calculate the false omission rate $\mathbf{FOR} = \frac{\mathbf{FN}}{\mathbf{FN+TN}}$. Now for the next sample have the cost function $\mathcal{C}(b_l, b_u)$ defined as follows:

$$\begin{aligned} & C_f(\mathbf{FOR}(b_l)) \cdot [\mathbf{TN}(b_u) + \mathbf{FN}(b_l)] + c_r \cdot \mathbf{FP}(b_u) + \\ & c_l \cdot (1 - p_o) \int_{b_l}^{b_u} g_g(x)\, dx + c_l \cdot p_o \int_{b_l}^{b_u} g_b(x)\, dx. \end{aligned} \tag{4}$$

This function is our minimisation target for which we choose b_l and b_u accordingly:

$$\begin{aligned} \min_{b_l, b_u} \quad & \mathcal{C}(b_l, b_u, g_b, g_g, p_0) \\ \text{s.t.} \quad & b_l \leq b_u \\ & b_l, b_u \in \overline{\mathbb{R}} \end{aligned} \tag{5}$$

where $\overline{\mathbb{R}}$ is the set of the extended real numbers which additionally contains plus and minus infinity, i.e. the union of sets $\mathbb{R} \cup \{-\infty, +\infty\}$.

We list the following possible solutions and explanations of their respective meanings for human inspection and machine inspection.

$\boldsymbol{b_l, b_u = -\infty}$ Every inspected sample should be rejected as an outlier. This will be in most cases a non-meaningful answer. However, in some cases where there is an extra station for rejected pieces, this would mean that every piece should go through this station. In some cases, we could have a very high probability for the outlier percentage p_0 temporally as a defective machine could produce parts which all are outliers. We will discuss the problem of non-independence of outliers in Section 3.7.

$\boldsymbol{b_l = -\infty,\ b_u \in \mathbb{R}}$ This means pieces with outlier score $> b_u$ should be rejected and any piece below this score should be inspected by a human. This further means that double-checking any piece by the outlier model and human is a necessity for passing on the processing. Items which are cheap to produce but have a costly role could qualify for such a solution.

$\boldsymbol{b_l, b_u \in \mathbb{R},\ b_l < b_u}$ There are three intervals. One where every piece will be accepted, one where every piece will be inspected by a human and accepted or rejected and one where every piece is rejected. Hence the machine will accept and reject some samples without human intervention while leaving the not-so-clear cases for the human.

$\boldsymbol{b_l, b_u \in \mathbb{R}, b_l = b_u}$ A piece will either be accepted by the machine or rejected. No human intervention is necessary for cost-optimal results.

$\boldsymbol{b_l \in \mathbb{R}, b_u = +\infty}$ Every piece with outlier score $> b_l$ should be inspected and with smaller b_l accepted. Rejected pieces could be very costly to discard on the one hand, on the other the outlier detection could have very high precision.

$\boldsymbol{b_l, b_u = +\infty}$ Every piece should be accepted, hence no inspection should be conducted and there is no benefit in running the outlier detection. Still later on the outlier probability p_0, as we will discuss in Section 3.7, could be changing. Hence there could be a benefit in running the outlier detection when the percentage of outliers is changing for the worse.

We may have multiple additional restrictions we can apply to the cost minimisation problem. In case of a very low outlier rate p_o, we may simplify the cost by setting $b_u = \infty$. The reason for this is that since there is no cut-off boundary b_u any extra burden of inspection of samples with very high outlier score will be very small compared to the inspection close to b_l, mainly dominated by the non-outlier. Moreover, when setting $b_u = \infty$ the optimisation problem becomes a single variable problem.

Often it will be the case that we have a fixed percentage of images, say p_f, which we can inspect due to, for instance, a fixed amount of available human labour. In this case, the following constraint will be added to the cost optimisation problem 5

$$p_f = (1 - p_o) \int_{b_l}^{b_u} g_g(x)\, dx + p_o \int_{b_l}^{b_u} g_b(x)\, dx. \tag{6}$$

If we additionally set $b_u = \infty$ as we discussed before we can already find the optimal b_l by just using this constraint. But even in this case, the considerations made before are still useful as we can still estimate the cost of our system and further estimate whether it is useful to employ or dismiss a human at a certain cost or estimate the cost saving for a higher or lower rate of inspection of samples.

Moreover, we can estimate costs of inspection done my m-many humans by taking multiple of p_f and hence get the following constraint to the optimisation problem

$$m \cdot p_f = (1 - p_o) \int_{b_l}^{b_u} g_g(x)\, dx + p_o \int_{b_l}^{b_u} g_b(x)\, dx. \tag{7}$$

Hence to find the optimal number of humans to employ for additional inspection we have to solve the optimisation problem for each possible m, where the lower bound is 0, i.e. no inspection, and the upper inspection limit is the amount of worker needed to inspect every possible sample produced. Similarly, we can also produce the exact amount, a possible non-integer amount, of workers needed by finding the solutions to the cost-optimisation problem without constraint and dividing by the inspection performance of an average human.

3.6 Cost-function for a calibrated output.

We now assume our model has probabilistic output, i.e. our model is calibrated. In this case, under some additional assumption for the cost of false-negative samples, the cost consideration and optimisations are drastically simplified. We could start to reconstruct the class-conditional distributions $P(X|Y)$ for the bad and good samples and hence get g_b and g_g and then apply the analysis as done before. This could be done by using the resulting distribution $P(Y|x) = M(x)$ and also inferring the distribution $P(X)$ from data, i.e. the distribution of an observed sample X having a certain value $[0, 1]$. But we can somewhat simplify the cost model by using the probabilistic nature of the classifier which resembles the conditional $P(Y|x) = M(x)$.

The marginal cost for an observed sample x with probability $M(x) = p$ being an outlier is as follows.

$$\mathbb{1}_{p<b_l} \cdot C_f(p) + c_r \cdot \mathbb{1}_{p>b_u} \cdot (1-p) + c_l \cdot \mathbb{1}_{b_l<p<b_u} \tag{8}$$

Note that we deviated from the equation $C_f(\mathbf{FOR})$ and apply C_f directly onto the value p of being an outlier. We can construct this new C_f from the original by constructing the distribution $P(X)$ and calculating the false omission rate by exploiting that $P(Y|x) = M(x)$ and the fact that there are only two classes, i.e. we use Bayes' Theorem to calculate $P(X > p|Y = 0)$ and further **FOR** in a similar fashion. With this, we can find a replacement function which takes p as input and resembles the original C_f.

We now look for the optimal parameters b_l and b_u. As a value for b_u, we choose the solution to the equation $c_r \cdot (1-x) = c_l$, if it is within the range of possible values. Hence we can find the upper boundary b_u of the human inspection interval as the solution to the following equation

$$\max\{0, \frac{c_r - c_l}{c_r}\}. \tag{9}$$

The value of b_l is determined by the solution of the equation $C_f(p) = c_l$ with the constraint that $0 \leq b_l \leq b_u$ holds. Hence in the case of a calibrated classifier and some additional assumptions, the estimation of outliers is simplified.

3.7 Non independence of outlier observations

In the case where we believe there is a non-independence of the series of observed data[4] we could for example increase the believed percentage of outliers p_o for the next few observed samples after observing an outlier. This ensures that the costs stay optimal for the next observed samples with higher anomaly probability. Note that in more complicated production environments we may observe pieces from multiple different machines. If possible one should keep track of all the machines involved, their maintenance status and other critical parameters to get a current assessment of the probability of outlier probabilities[27].

In lean manufacturing systems[5] there is a drive towards zero defects in each production step. If such a goal would be achievable, this would limit the amount of data to consider for estimating the outlier probability tremendously as only the last step would have to be considered. Still, zero-defects or even lower goals as six sigma have their limitations in the real world[19][18] and in a sufficiently complicated production step there will still be enough complexity to link the probability of defects to other parameters measured. As in reality often a bathtub-like curve of failures is observed[8][22]. In the middle of the bathtub curve, we have mostly random failures and near-constant error rates[22]. In this case, an independent assumption for the observation of outliers is justified. Moreover, the bottom of the bathtub curve can be prolonged with predictive maintenance[7] [27] and hence in such a hopefully long-lasting system state, it

[4] A broken machine could for example produce a sudden stream of defective parts.

is justified, as initially assumed, that the non-independence of outliers is a valid assumption for the whole life-cycle of our machine-learning model. Moreover, such a system as we described for outlier detection containing a probability model becomes itself a strong data source which then can be incorporated into a predictive maintenance scheme, as the observed sequence of good and bad products together with the model are a great help in identifying anomalous behaviour such as an increased observation of outliers.

4 Algorithm

In this section, we combine the observations established in the previous section into a combined algorithm. We note this algorithm in pseudo-code. For this see Algorithm 1 further below. As an input to our algorithm, we provide a one-class classification model M that needs N-many samples for initial training and testing. Further, we have some belief about the percentage of outliers p_o in the samples to be observed. Additionally, we have the positive monotone cost function C_f for false negative samples. Then there is a real positive value c_r representing the cost of a false positive sample. We also have some real positive value c_l which represents the cost of human labour for the labelling of a sample. Moreover, we fix a number of outliers we want to observe L. In summary, the algorithm then starts by letting humans label samples until we obtain a set D containing N-many samples labelled as good, i.e. samples with the label 0. We then use this data set D to train our classification model and obtain some trained model M_D. This model is then used to produce outlier scores for the following samples. Moreover, we produce scores for the test data split of D. The model is then used to find anomalous samples efficiently. We do this to obtain a probabilistic model by inferring a Gaussian curve of the good and a Gaussian curve of the bad samples or to update our belief about the prior distribution of the good and bad samples. Based on this we are finally able to find the cost optimal parameters b_l and b_u by solving the Optimisation Problem 5. This interval $[b_l, b_u]$ marks the outlier score where in-between human inspection will take place.

Note that furthermore, we can replace Step 3 to Step 10 of the Algorithm 1 with an active learning algorithm. Especially, if labelling can be done asynchronously the labelling cost should not increase with this approach. This would also improve the performance of the model as our model will potentially see a more diverse input of samples. For this, we need some one-class active learning algorithm which can be found in [24] for example. Moreover, the updated data set D of step 21 could also be used to improve the model M. Especially using the outlier samples could be beneficial for the performance and active learning part by training another one-class model for these samples[21].

Algorithm 1 (Find optimal interval for human inspection)**.**

```
1: initialisation: $p_o, C_f, c_r, c_l, N, L$
2: $n \leftarrow 0$
3: for $n < N$ do
4:     wait for next sample $s$
```

```
    get label $l(s)$ (by human)
    $n \leftarrow n + 1 - l(s)$
    $p_o \leftarrow$ belief update through observed $l(s)$
end for
return training data set $D$ , $p_0$
$M_D \leftarrow$train model with $D$
$b'_l \leftarrow$ (see Section 3.4 for possible computations)
$k \leftarrow 0$
for $k < L$ do
    get next sample $s$
    if $b_l < M_D(s)$ then
        get label $l(s)$ (by human)
    end if
    $k \leftarrow k + l(s)$
    $p_o \leftarrow$ belief update through observed $l(s)$
end for
return updated data set $D$, $p_0$
$g_g, g_b \leftarrow$ interfere Gaussian from data $D$
solve $\min_{b_l,b_u} \mathcal{C}(b_l, b_u, g_b, g_g, p_0)$
Ensure: Model $M_D$ and inspection interval values $b_l, b_u$
```

5 Experiments

We conduct the D'Agostino's K2 normality test [4] on the test data of MVtec by using the implementation in the SciPY-package[26]. We remind the reader that the normality test is a statistical test with the null hypothesis being that the given data is Gaussian distributed. We use this test for every test data set of each MVtec data set separately on the outlier samples and good samples with a p-value of 5%. Only for the *wood* data set the null hypothesis cannot be dismissed. One problem encountered often is that the Patchcore algorithm only produces a positive outlier score but the computed Gaussian has negative values which leads to a very low p-value. Here another distribution which only takes positive values could be more feasible, such as the truncated Gaussian distribution. The p-values of the normality test for the *wood* data set are the following. We have 93.7% for the outlier samples and a p-value of 47.1% for the good samples. A histogram and the fitted Gaussians can be seen in Figure 6, where we took 60 outlier samples and 19 good samples in this test data set.

We also computed the AUC-scores for the ROC curves of the Patchcore model outlier scores on the wood test data with the help of scikit-learn[16]. Additionally, with the same software we computed the AUC-scores for millions of samples from the two fitted Gaussian distributions of these data sets with the same ratio of outliers and good samples. The AUC-score for the former data set is 99.04% and that of the latter data set is 99.09%.

Due to the high p-values of the normality tests and the similar AUC-scores of the test data and the fitted Gaussian distributions, we conclude that the model is a good fit for the test data's outlier scores of the Patch-

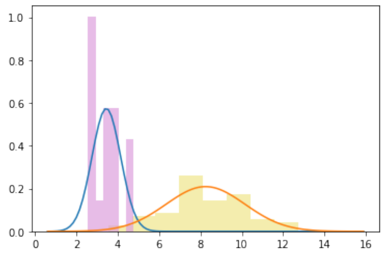


Figure 6: Histograms of the outlier scores of the *wood* data set of the MVtec data set with fitted Gaussian distributions. The left curve and histogram represent the good samples and the right curve and histogram represent the bad samples.

core model on the *wood* data set.

We also did a small computational study with the cost model we introduced (see Equation 4). We now search for some examples to verify that indeed there are non-trivial solutions to the cost model. For that, we use the Gaussian distributions fitted to the wood data set together with some cost parameters and apply the proposed cost model. We compare the optimal solution of the cost equation and the boundary for optimal F1 score for the test data which is without human intervention. For that, we assume that our production line produces 1% outliers, and our labour cost c_l for inspection is normalised to 1. As a cost function C_f we use the following $C_f = \alpha \cdot (\exp(\beta \cdot \text{FOR}) - 1)$. This function ensures that if the false omission rate is 0 we have no cost associated with this function. The lost revenue of a good item discarded is c_r which we let range as integers from 0 to 40. The parameter α we let range from 1 to 10 with a step of 1 and the parameter β from 5 to 200 with a step of 5, both as integers. Note that a value of $\alpha = 1$ and $\beta = 200$ would lead to a cost of e^2 for every item passing the inspection. Hence if $c_r \leq 7$ is in our search this would mean that the product is essentially not sellable. We search for minimal solutions with the downhill simplex algorithm in SciPy by starting at $b_l = 3.5$ and $b_u = 4$ with the sub-condition that $0 < b_l < b_u$. Although this does not necessarily find the global minima we can still infer from the local minima that it is worth considering a human-in-the-loop approach. In particular, we find for more than 90% of the search space a minimal solution with an average cost improvement over the F1-optimal cut-off value of 57%. Most values of b_l and b_u range between 3 and 7. This implies that if the cost structure of the inspection task and results are known and correct, it is well worth finding an optimal solution to our cost model.

6 Discussion and future work

We established a theory for the cost-optimal selection of samples for additional human labelling of one-class classification models. For this, we established a cost model and showed how to infer probabilistic knowledge of the samples online and offline to establish a cost-optimal decision for a human inspection boundary in the outlier score. Moreover, we have merged this into an algorithm which can be applied in production. Further, we conducted a small experimental study checking the MVtec data sets "normality" and found that most data sets have values close to 0 while being positive to be of Gaussian distribution. Moreover, we searched for minima for the cost equation in the parameter space and more than halved the cost compared to F1 optimisations.

For now, we have not considered the case of retraining the model and we can assume that this will be done occasionally until the economic evaluation stabilises or the performance is satisfactory. Also, the problem of a timely dependence on the occurrence of outliers, which could stem from faulty machines and certain maintenance cycles, was discussed. At worst there could be no outlier samples or only a very biased selection of them. A detailed analysis of the practical relevance of this problem and a field study is an interesting topic for future investigation. There could also be potential for future work, especially in the case where the one-class problem is a moving target, i.e. the golden sample changes over time. The case for selecting valuable examples for improving the model performance also seems an interesting area not yet considered and will probably require an extra model which is also trained with the outliers. Another not yet-used feature is utilising the presentation of anomalous areas on the image for better outlier visualisation for the user decision. There, another optimisation problem arises which is the optimisation of the cut-off parameter for the selection of the anomalous area. A more general question is the question of an appropriate visualisation to improve human performance. From the general perspective, the work looked into combining machines with human labour to better (in terms of cost) conduct classification problems with regard to industrial applications.

References

[1] Paul Bergmann et al. "MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9592–9600.

[2] Paul Bergmann et al. "The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection". In: *International Journal of Computer Vision* 129.4 (2021), pp. 1038–1059.

[3] Chengliang Chai et al. "Human-in-the-loop outlier detection". In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2020, pp. 19–33.

[4] RALPH D'AGOSTINO and E. S. PEARSON. "Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{b1}$". In: *Biometrika* 60.3 (Dec. 1973), pp. 613–622.

ISSN: 0006-3444. DOI: 10.1093/biomet/60.3.613. eprint: https://academic.oup.com/biomet/article-pdf/60/3/613/576953/60-3-613.pdf. URL: https://doi.org/10.1093/biomet/60.3.613.

[5] Uwe Dombrowski and Tim Mielke. *Ganzheitliche Produktionssysteme: Aktueller Stand und zukünftige Entwicklungen*. Jan. 2015. ISBN: 978-3-662-46163-1. DOI: 10.1007/978-3-662-46164-8.

[6] Chuan Guo et al. "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

[7] Hashem M Hashemian. "State-of-the-art predictive maintenance techniques". In: *IEEE Transactions on Instrumentation and measurement* 60.1 (2010), pp. 226–236.

[8] Urban Hjorth. "A Reliability Distribution With Increasing, Decreasing, Constant and Bathtub-Shaped Failure Rates". In: *Technometrics* 22.1 (1980), pp. 99–107. DOI: 10.1080/00401706.1980.10486106. eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1980.10486106. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1980.10486106.

[9] Baihong Jin et al. "A one-class support vector machine calibration method for time series change point detection". In: *2019 IEEE International conference on prognostics and health management (ICPHM)*. IEEE. 2019, pp. 1–5.

[10] Parimal Kopardekar, Anil Mital, and Sam Anand. "Manual, hybrid and automated inspection literature and current research". In: *Integrated Manufacturing Systems* 4.1 (1993), pp. 18–29.

[11] Anil Mital, M Govindaraju, and B Subramani. "A comparison between manual and hybrid methods in parts inspection". In: *Integrated Manufacturing Systems* 9.6 (1998), pp. 344–349.

[12] Eduardo Mosqueira-Rey, David Alonso-Rios, and Andres Baamonde-Lozano. "Integrating Iterative Machine Teaching and Active Learning into the Machine Learning Loop". In: *Procedia Computer Science* 192 (2021), pp. 553–562.

[13] Robert Munro. *Human-in-the-loop machine learning*. New York, NY: Manning Publications, Oct. 2021.

[14] Kevin Murphy. *Conjugate Bayesian analysis of the Gaussian distribution*. Nov. 2007. URL: https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf.

[15] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. "Obtaining well calibrated probabilities using bayesian binning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 29. 1. 2015.

[16] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[17] Pramuditha Perera, Poojan Oza, and Vishal M Patel. "One-class classification: A survey". In: *arXiv preprint arXiv:2101.03064* (2021).

[18] Foivos Psarommatis et al. "Zero defect manufacturing: state-of-the-art review, shortcomings and future directions in research". In: *International journal of production research* 58.1 (2020), pp. 1–17.

[19] John S Ramberg. "Six sigma: Fad or fundamental". In: *Quality Digest* 6.5 (2000), pp. 30–1.

[20] Karsten Roth et al. "Towards total recall in industrial anomaly detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14318–14328.

[21] Patrick Schlachter and Bin Yang. "Active learning for one-class classification using two one-class classifiers". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pp. 1197–1201.

[22] Abs Shagluf, Andrew Longstaff, and Simon Fletcher. "Maintenance Strategies to Reduce Downtime Due to Machine Positional Errors". In: Sept. 2014.

[23] Albert Thomas, Vincent Feuillard, and Alexandre Gramfort. "Calibration of One-Class SVM for MV set estimation". In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2015, pp. 1–9. DOI: `10.1109/DSAA.2015.7344789`.

[24] Holger Trittenbach, Adrian Englhardt, and Klemens Böhm. "An overview and a benchmark of active learning for outlier detection with one-class classifiers". In: *Expert Systems with Applications* 168 (2021), p. 114372.

[25] Juozas Vaicenavicius et al. "Evaluating model calibration in classification". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, Apr. 2019, pp. 3459–3467. URL: `https://proceedings.mlr.press/v89/vaicenavicius19a.html`.

[26] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: `10.1038/s41592-019-0686-2`.

[27] Tiago Zonta et al. “Predictive maintenance in the Industry 4.0: A systematic literature review”. In: *Computers & Industrial Engineering* 150 (2020), p. 106889.