International Journal of Production Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tprs20

Few-shot learning for defect detection in manufacturing

Patrik Zajec, Jože M. Rožanec, Spyros Theodoropoulos, Mihail Fontul, Erik Koehorst, Blaž Fortuna & Dunja Mladenić

To cite this article: Patrik Zajec, Jože M. Rožanec, Spyros Theodoropoulos, Mihail Fontul, Erik Koehorst, Blaž Fortuna & Dunja Mladenić (27 Feb 2024): Few-shot learning for defect detection in manufacturing, International Journal of Production Research, DOI: 10.1080/00207543.2024.2316279

To link to this article: https://doi.org/10.1080/00207543.2024.2316279

Published online: 27 Feb 2024.

INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
https://doi.org/10.1080/00207543.2024.2316279

Taylor & Francis
Taylor & Francis Group

Few-shot learning for defect detection in manufacturing

Patrik Zajec[a,b,*], Jože M. Rožanec[a,b,c,*], Spyros Theodoropoulos[d,e], Mihail Fontul[f], Erik Koehorst[g], Blaž Fortuna[b,c] and Dunja Mladenić[b]

[a]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia; [b]Jožef Stefan Institute, Ljubljana, Slovenia; [c]Qlector d.o.o., Ljubljana, Slovenia; [d]Department of Digital Systems, University of Piraeus, Piraeus, Greece; [e]Department of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece; [f]Iber-Oleff S.A., Coimbra, Portugal; [g] Philips Consumer Lifestyle BV, Drachten, The Netherlands

ABSTRACT

Quality control is being increasingly automatised in the context of Industry 4.0. Its automatisation reduces inspection times and ensures the same criteria are used to evaluate all products. One of the challenges when developing supervised machine learning models is the availability of labelled data. Few-shot learning promises to be able to learn from few samples and, therefore, reduce the labelling effort. In this work, we combine this approach with unsupervised methods that learn anomaly maps on unlabelled data, providing additional information to the model and enhancing the classification models' discriminative capability. Our results show that the few-shot learning models achieve competitive results compared to those trained in a classical supervised classification setting. Furthermore, we develop novel active learning data sampling strategies to label an initial support set. The results show that using sampling strategies to create and label the initial support set yields better results than selecting samples at random. We performed the experiments on four datasets considering real-world data provided by *Philips Consumer Lifestyle BV* and *Iber-Oleff - Componentes Tecnicos Em Plástico, S.A.*

ARTICLE HISTORY
Received 20 February 2023
Accepted 22 January 2024

KEYWORDS
Smart manufacturing; quality inspection; automated visual inspection; artificial intelligence; active learning; few-shot learning

1. Introduction

The increasing connectivity capabilities and adoption of digital technologies have enabled the digitalisation of manufacturing and originated new manufacturing paradigms known as Industry 4.0 and Industry 5.0 (Benbarrad et al. 2021; Lenka, Parida, and Wincent 2017). While Industry 4.0 is concerned with leveraging new technologies (e.g. the internet of things, cloud computing, and artificial intelligence, among others) to increase productivity across the value chain and enable the efficient production of goods (Lim, Zheng, and Chen 2020), Industry 5.0 is concerned with how such technologies can be used and applied to achieve a human-centric workspace, and thus changing the role of the operator (EC2 n.d.; Lied, Mogos, and Powell 2020; Nahavandi 2019).

Quality control is a key phase of the manufacturing process, ensuring the products conform to specific requirements and specifications (Yang et al. 2020) and therefore is a precondition to building a brand's reputation, build trust with the consumer, and loyalty. While such inspection has been frequently performed manually, there is an increasing trend to automate the quality inspection process. Some of the advantages of such automation are increased scalability (Chin and Harlow 1982; Chouchene et al. 2020), homogeneous defect inspection criteria (See 2012), and the ability to trace defects' root causes to solve issues in the production process proactively. Research shows that Industry 4.0 technologies applied to defect inspection have the potential to realise a substantial increase in productivity Tortorella et al. (2023). Automated visual quality inspection is one such approach (Abd Al Rahman and Mousavi 2020).

Many approaches have been tried to build automated visual inspection models. Most recent approaches leverage advances in machine learning to determine whether a defect exists and eventually determine the type of defect. While unsupervised approaches detect whether a manufactured piece is defective, they do not provide information on the defect detected. Supervised methods can provide such information, which requires manually labelling samples of good and defective manufactured

CONTACT Jože M. Rožanec joze.rozanec@ijs.si Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana 1000, Slovenia Jožef Stefan Institute, Jamova 39, Ljubljana 1000, Slovenia Qlector d.o.o., Rovšnikova 7, Ljubljana 1000, Slovenia

*Jože M. Rožanec and Patrik Zajec are co-first authors with equal contribution and importance pieces. Data labelling is a costly operation. While certain approaches (e.g. active learning) can alleviate the labelling effort, annotating a few hundred images per defect is usually necessary to ensure the machine learning model learns appropriately. Few-shot learning is a recent approach that aims to reduce the number of labelled instances required to train a classifier. Using such an approach, we extend the experiments performed in our work described in '*Towards a Comprehensive Visual Quality Inspection for Industry 4.0*' (Rožanec, Zajec, Trajkova et al. 2022).

This research aims to determine how few-shot learning can be used and enhanced in the context of defect detection, ensuring the least amount of data is used while maximising the models' discriminative performance. This would reduce the effort required to develop new defect detection models and the time to train them, increasing agility while reducing development costs in manufacturing and conforming to the required quality levels for multiple products. In particular, the goals pursued were:

- based on the findings by Rožanec, Zajec, Theodoropoulos et al. (2022), assess how combining images and DRAEM (Discriminatively trained Reconstruction Anomaly Embedding Model) (Zavrtanik, Kristan, and Skočaj 2021) anomaly maps (which signal potential defects) enhances the classification quality and generalises to few-shot learning scenarios;
- contrast traditional supervised learning with an artificially induced imbalance with few-shot learning to assess the effectiveness and shortcomings of both approaches;
- study how dataset impurity levels affect DRAEM models (to avoid labelling data, which would defeat the few-shot learning purpose);
- develop novel active learning strategies that can assist in creating better support sets

The main innovation points of this research are:

- the use of the information extracted from unsupervised methods to enhance supervised few-shot classification learning and performance;
- the development of a novel active learning strategy that leverages explainable artificial intelligence insights for data selection

The experiments were performed with real-world data provided by *Philips Consumer Lifestyle BV* and *Iber-Oleff - Componentes Tecnicos Em Plástico, S.A.*. The machine learning models were evaluated with the AUC ROC metric to inform the discriminative power of the machine learning models.

This paper is organised as follows. First, Section 2 describes related work, Section 3 describes the *Philips Consumer Lifestyle BV* and *Iber-Oleff - Componentes Tecnicos Em Pláistico, S.A.* use cases and datasets. Section 4 describes the experiments we performed, and Section 5 informs the results we obtained. Finally, Section 7 concludes and describes future work.

2. Related work

2.1. Automated visual inspection

Traditional visual inspection involves human operators inspecting the manufactured pieces to determine whether they are defective. Many drawbacks to this approach have been observed. Among them, manufacturing companies are concerned about the limited scalability of the approach (e.g. given that human inspectors can work for a limited amount of time, and the resources required to train new human inspectors usually grow proportionally to the production scale). Furthermore, the inspection performed by each human inspector is subjective. Therefore, an inherent inspector-to-inspector inconsistency exists regardless of the human inspectors' proficiency in the process. Such discrepancies are influenced and magnified by factors related to the task (e.g. defect rate, the complexity of visual inspection), the inspector (e.g. visual acuity or experience), the environment in which the inspection is performed (e.g. lighting, shift duration and time of the day), and organisational (e.g. management support or incentives) and social aspects (e.g. isolation or opportunity for consultation) (Cullinane et al. 2013; Kujawińska, Vogt, and Hamrol 2016; See 2012).

2.1.1. Background

The automated visual inspection aims to address the abovementioned issues. It guarantees scalability by creating software capable of inspecting manufactured products and determining whether they are defective. Furthermore, inspector-to-inspector inconsistency is eliminated, given a single criterion for product quality is established. An automated visual inspection enables non-destructive testing in quality control to identify functional and cosmetic defects (Chin and Harlow 1982). Cameras provide visual input, which can be processed with different techniques (Czimmermann et al. 2020). State-of-the-art (SOTA) automated visual inspection techniques leverage deep learning techniques (Aggour et al. 2019; Pouyanfar et al. 2018), which have demonstrated super-human performance on many machine vision tasks (O'Mahony et al. 2020). Such models can be either supervised or unsupervised. The unsupervised methods allow for the discrimination of defective manufactured pieces without any labelled data. While such an approach is attractive given that no data labelling is required, it does not provide information on the defect type and is, therefore, unsuitable for all manufacturing processes. On the other hand, supervised models can discriminate between different types of defects and, therefore, can be helpful in production when different levels of quality must be satisfied. For example, some imperfections can be cosmetic, while others may affect product functionalities. Therefore, different thresholds can be used for them. Furthermore, information on the type of defect can be used in many settings to determine the root causes of such defects and take appropriate action. Nevertheless, supervised models require data labelling, which is a time-consuming and error-prone task that must be performed by humans (Y. Wang et al. 2018).

Multiple artificial intelligence approaches have been researched to reduce the labelling effort. One such approach is the active learning paradigm, which assumes a constrained capacity to provide learning samples to a machine learning model and that the learning process can be improved by carefully selecting the data instances to maximise learning towards a given objective (Settles 2009). Such data instances can be either sampled from actual data or artificially generated. A second paradigm is transfer learning, which aims to transfer knowledge acquired from another source or domain where data is abundant and apply it in a different setting where data regarding the origin or domain is scarce. Furthermore, domain adaptation is a variation of this approach, where the source and target tasks are the same, but the source and target domains differ. Fourth, meta-learning aims to learn meta-knowledge across tasks and apply it to a concrete task based on task-specific information. Finally, few-shot learning compensates for the lack of supervised data by reframing the classification problem and learning how close the data instances between classes are. Furthermore, it leverages the lack of data using meta-learning, generating synthetic samples, or recurring to transfer learning (using a data representation learned on a different dataset and training a new classifier) (Parnami and Lee 2022).

2.1.2. Machine learning approaches to automated visual inspection

There have been many works from various industrial sectors on the automation of visual quality inspection relying on machine learning and deep learning methods. For instance, in an early example of an inspection of Printed Circuit Boards (Duan et al. 2012), statistical shape models micro-drill bit defects were combined with dimensionality reduction techniques (Principal Component Analysis and Linear Discriminant Analysis) to create input features for various models, including Support Vector Machines and shallow Multi-layer perceptrons. The promising results of Support Vector Machines on custom extracted features were identified even earlier in the inspection of rolled steel (Jia et al. 2004), which managed to integrate them in a fast (six seconds per 1MB image) real-time system. More recently, Support Vector Machines and genetic algorithms were successfully used to detect porosity defects in the welding process of aluminum by combining extracted features from various sources such as spectral and X-ray data (Huang et al. 2017). Gobert et al. (2018) examined the metallic power bed fusion process in additive manufacturing also through the SVM-based classification of features originating from a digital single-lens reflex (DSLR) camera and labelled in a semi-automatic way with the help of CT scans. Despite the success of methods based on custom feature extraction combined with a traditional machine learning classifier (such as SVMs), later approaches use deep learning, especially Convolutional Neural Networks (CNNs), which operate directly on images, adaptively extracting features during their training process. While many pre-trained CNNs on large datasets can be used off the shelf and finetuned to a specific use case, Villalba-Diez et al. (2019) found it more advantageous to train a custom shallow CNN from scratch, specifically tailored to their Printing Industry use case. What appeared challenging to them was the standardisation of input image conditions, especially regarding controlling image brightness. Liqun, Jiansheng, and Dingjin (2020), on the other hand, followed the path of transfer learning and found that the classification of vehicle parts via finetuning a pre-trained VGG16 model produced higher accuracy in comparison to a Support Vector Machine over Histogram of Gradients (HoG) features. It could well be the case that more complex products such as vehicle parts need these complicated but versatile conditions to conform to modern inspection application requirements such as scalability, agnosticity to different inputs, and quick retraining process as outlined by Chouchene et al. (2020). Yu et al. (2023) proposed the Cascaded Adaptive Global Location Network. This novel deep neural network combines residual, feature pyramid, and cascade adaptive tree-structure region proposal networks for feature extraction and uses a global localisation regression to perform defect detection. The authors applied the model to defect detection on steel surfaces. Zhao et al. (2023) proposed a multi-surface defect detection method that performs region segmentation, feature extraction, and defect detection, enabling an efficient quality control of universal joint bearings.

Beltrán-González, Bustreo, and Del Bue (2020) successfully combined CNNs with Long Short-Term Memory Networks (LSTMs) to identify the presence of debris in avionic component ducts. Finally, Shahin et al. (2023) reported how the YOLO v7 model was used to discriminate defective packages and prevent them from moving into shipping operations. For detailed systematic literature reviews on this field, we encourage the authors to read the excellent works by Ren et al. (2022), Konstantinidis et al. (2023), and Abd Al Rahman and Mousavi (2020).

2.2. Few-shot learning

Few-shot learning is a machine learning approach where the learner aims to acquire experience to solve a specific task with only a few data samples. As for any machine learning approach, the success of such learning is measured with a particular metric suitable to the specific goal (Y. Wang et al. 2020).

2.2.1. Background

When considering how previous experience is deemed to enable learning from a few data instances, few-shot learning approaches can either adjust the data (e.g. augment the data set with samples from other datasets or use unlabelled data), the model (e.g. acquire knowledge on another dataset), or algorithm (e.g. adapt hyperparameters based on prior meta learned knowledge) (Y. Wang et al. 2020).

Parnami and Lee (2022) categorise few-shot learning approaches into meta-learning-based and non-meta-learning-based few-shot learning approaches. The non-meta-learning-based approaches consider few-shot learning approaches derived from transfer learning. On the other hand, meta-learning-based approaches are divided into two categories: *hybrid* and *main* approaches. Among the *main* approaches, we find metric-based, optimisation-based, and model-based meta-learning. In a classification setting, metric-based approaches attempt to learn a mapping from input data to an embedding space, ensuring that data instances from the same class remain close to each other and distant from different classes. Therefore, the distance to the nearest neighbours can be used to determine the class of a particular instance in test time. Optimisation-based few-shot learning techniques aim to optimise the limited training data while still achieving good generalisation. They usually do so by learning in two stages: a task-specific learner is used to solve a specific task, and a non-task-specific meta-learner is used to learn from the experience acquired through multiple tasks and direct further learning. In episodic training, the meta-learner updates the learner model's parameters based on the experience acquired through the many tasks it trained on. Finally, model-based meta-learning does not make any assumptions on priors but focuses on architectures tailored for fast learning. Among such architectures, we find memory-based architectures (Cai et al. 2018), rapid-adaptation architectures (Munkhdalai and Yu 2017), and other approaches (Mishra et al. 2017).

When dealing with classification, few-shot learning tries to compensate for the lack of data by framing the learning problem to learn similarities and differences between classes. This approach is fundamentally different from traditional machine learning approaches, where the algorithm is trained to learn what constitutes a particular class. While the classification outcomes are the same (decide whether a data instance corresponds to a given class), the learning process is not. Furthermore, few-shot learning classification requires a slightly different training setup. The train set (a.k.a. support set) comprises $M * K$ data instances corresponding to M classes and K examples per class. The classes present in the support set are usually referred to as *base classes*. Furthermore, a query set contains the images to be classified, which can correspond to *base classes* and *novel classes* (not seen in the support set). Nevertheless, this task definition does not consider class-imbalance scenarios, which are frequent in the real world. How to mitigate performance drops where a class imbalance is present remains an open challenge (Ochal et al. 2021).

2.2.2. Few-shot learning for automated visual inspection

The advantage of few-shot learning has made it an interesting approach when developing machine learning models for automated visual inspection. Lv and Song (2019) developed a few-shot learning approach to detect defects on bar surfaces. The model involved a convolutional neural network (CNN) in extracting image features and a relation network to compute a similarity score between pairs of images. The authors used a Squeeze-and-Excitation Network as an attention module to enhance features describing defects. They employed Mean-Pooling to preserve background information and distinguish better between pseudo and real defects. A similar approach was developed later by Takimoto et al. (2022), who performed anomaly detection using a convolutional neural and Siamese network with an attention mechanism to detect defects on the MVTec dataset. They proposed using a pair-balanced contrastive loss to account for the effect of data imbalance. Furthermore, the attention mechanism aimed to increase the distance between data instances of different classes in the embedding space. The Siamese network was used to perform metric learning and learn to discriminate defective and non-defective products based on the learned metrics. H. Wang, Li, and Wang (2021) proposed an incremental few-shot learning framework and executed experiments using the Faster R-CNN as a backbone model. The authors aimed to detect defects on steel surfaces. To enhance the model's training performance, the authors considered a diverse set of input images must be used and, therefore, resorted to performing data augmentation to guarantee such diversity by applying image transformations. Wu et al. (2021) described a few-shot learning approach for defect detection in lithium batteries. In particular, they considered exposure fusion to capture batteries' reflectivity and convey 3D information in a 2D image. Furthermore, they used data augmentation to enrich the datasets and label propagation to overcome the shortage of labelled data. The few-shot learning model was based on a ResNet-10 feature extractor and a fully connected layer to perform classification. Furthermore, Zhan, Zhou, and Xu (2022) described how prototypical networks were used to perform automated fabric defect classification. Furthermore, the authors used class activation mapping to visualise and interpret the regions relevant to the classification of a particular defect class. Zhang et al. (2020) described using few-shot with model-agnostic meta-learning to detect defects on bearings. The implementation considered convolutional neural networks treating the identification of the various types of defects as different tasks and then a meta-learner to learn the best parameters across the classification tasks to accelerate learning. Xu and Ma (2022) applied few-shot learning for auto parts defect detection. The authors compared the ProtoNet, the FEAT, and a custom network based on the ProtoNet and ECA-Net with an attention mechanism.

2.3. *Research gap*

While few-shot learning has been applied to defect detection, research has been mainly focused on developing novel deep-learning architectures that would issue better classification results. Furthermore, little research inquired into how few-shot learning models can benefit from carefully selected samples used to train each episode and input data enriched with cues about possible defects. This research aims to bridge this gap while researching active learning strategies that consider explainable artificial intelligence insights to select relevant data instances. In the context of current research, our studied approaches follow the trend towards data-centric instead of model-centric solutions. As explained in Singh (2023), as models gain in sophistication and their implementations become readily availably through different machine learning libraries, the predictive performance returns on model optimisation diminish, while the costs of developing and improving such models increase both regarding development effort and computational resources. This has led researchers to seek more impactful improvements in techniques that improve the quality or saliency of the input data. The use of heatmap-enhanced image inputs and active learning follows this trend in the context of few-shot visual quality inspection.

3. Use cases and datasets

We performed the experiments on real-world data provided by *Philips Consumer Lifestyle BV* (The Netherlands) and *Iber-Oleff - Componentes Tecnicos Em Plástico, S.A.* (Portugal). *Philips Consumer Lifestyle BV* manufacturing plant in Drachten is considered one of Europe's most important Philips development centres and produces many household appliances. The three datasets provided by them correspond to different products: (a) logo prints on shavers (see Figure 1), (b) deco cap (covers the centre of the metal shaving head and leaves room for a print to identify it from other types - see Figure 2), and (c) shaft (toothbrush part that transfers the motion from the handle to the actual brush - see Figure 3).

The *shavers* dataset contains 3.518 images with the heaviest imbalance among the datasets (the defective products account for almost 24% of the dataset). Two defects were labelled: double-printed logos and those with interrupted printing. The *deco cap* dataset contained 592 images and was labelled for two imperfections: flowlines and marks. The defects account for almost two-thirds of the dataset. Finally, the *shaft* dataset has

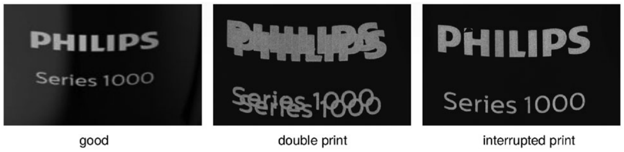

Figure 1. Sample from the *Philips Consumer Lifestyle BV* shavers dataset.

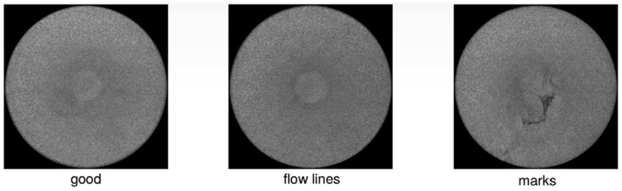

Figure 2. Sample from the *Philips Consumer Lifestyle BV* deco cap dataset.

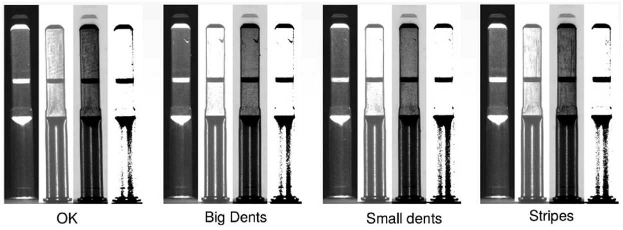

Figure 3. Sample from the *Philips Consumer Lifestyle BV* shaft dataset.

Table 1. Datasets description, describing label types and the number of data instances per label for each dataset.

Dataset	Class	Number of examples
Deco cap	good	203
	flowlines	198
	marks	191
Shaft	good	528
	big dents	2616
	small dents	954
	stripes	151
Shavers	good	2676
	double printing	244
	interrupted printing	598
IBER	good	2114
	failure 1	39
	failure 2	4
	failure 3	9

4.249 images and was labelled for three kinds of defects: big dents, small dents, and stripes. The images regarding defective items account for 38% of the dataset's images. We provide a more detailed description of the datasets in Table 1. Regardless of the product inspected, the manual inspection of the abovementioned products requires inspectors to spend several seconds handling and inspecting the product and determining whether it is defective.

Iber-Oleff - Componentes Tecnicos Em Plástico, S.A., on the other side, provided a dataset (we named it *IBER*) related to automobile air vents they manufacture. The air vents have three components of interest: housing, lamellas (used to direct the air), and plastic links (which keep the lamellas tied together). A visual inspection is performed to determine whether (a) the fork is leaning against the support and correctly positioned, (b) the plastic link is present, (c) the lamella 1 is present, and the link is correctly assembled, and (d) the lamella 3 is present, and the link is correctly assembled. We describe the datasets in detail in Table 1.

Among the expected benefits of automating the visual inspection are savings regarding manual work, increased process scalability, and assurance that the same criteria are used to determine whether a product is faulty. Furthermore, this research aims to provide insights enabling a solution requiring few labelled samples to train a machine learning model while satisfying the required product quality levels. By doing so, the labelling effort is minimised, and greater flexibility is provided to the manufacturing plant to address the visual quality inspection of other existing and new products.

4. Experiments

For this research, we conducted a series of experiments (see Table 2) to understand how few-shot learning could be applied to visual inspection with two purposes: (a) automating the visual inspection of manufactured products and (b) minimising data labelling requirements. We performed four experiments, with the following objectives: (i) compare multiple few-shot learning approaches on the given image datasets, (ii) understand whether anomaly maps can help machine learning models learn better, (iii) find out whether a particular unsupervised technique used to create anomaly maps can be trained on all data (images corresponding to good and defective products) without degrading the classifiers' performance (and the resulting quality of anomaly maps), and (iv) compare multiple active learning techniques where the support set is not selected randomly but following some active learning technique and whether this can lead to better results.

Table 2. Brief description of experiments performed: their aim, method utilised, and relevant metrics.

Experiment	Aim	Method	Metrics
1	Compare the performance of few-shot learning approaches applied to our use cases.	see Table 3	AUC ROC
2	Understand whether using DRAEM anomaly maps results in better performance of the few-shot learning approaches.	see Table 3	AUC ROC
3	Understand how data impurity affects DRAEM and how the resulting anomaly maps impact performance.	Training DRAEM model on dataset with different impurity levels and measuring the performance change of ViT backbone pre-trained with DINO and no meta-training as few-shot model.	AUC ROC
4	Compare active learning techniques used for selecting the support set in few-shot learning and determine how they impact few-shot performance	Selecting the support set using different active learning strategies and compare the performance change of the ViT backbone pre-trained with DINO and no meta-training.	AUC ROC

The few-shot learning models were trained considering one or five labelled images per class and the few-shot learning pipeline proposed by Hu et al. (2022a), which consists of three steps: pre-training, meta-training, and fine-tuning. The pre-training stage is devoted to training a backbone model used as a feature extractor in a few-shot learning setting. We did not perform the training but opted for pre-trained models on the ImageNet dataset (Russakovsky et al. 2015) either using the cross-entropy loss in the supervised setting or using the self-supervised DINO (Caron et al. 2021) objective. We sometimes performed meta-training on the Meta-Dataset (Triantafillou et al. 2019) or MVTec-Capsule (Bergmann et al. 2019). Fine-tuning was performed on the support set. We considered the ResNet-50 (He et al. 2016) and Vision Transformer (ViT) (Dosovitskiy et al. 2021) as backbone models, and ProtoNet (Snell, Swersky, and Zemel 2017) as a few-shot learner. Furthermore, we also considered the model described in Takimoto et al. (2022), which consisted of a Siamese network as a backbone, pre-trained on ImageNet with a cross-entropy loss and meta-trained on MVTec-Capsule; and a supervised classification model used for defect detection and described in Rožanec, Zajec, Theodoropoulos et al. (2022). We provide a detailed description of the models in Table 3.

The same test set was used across all the experiments, to ensure the results across experiments are comparable. We measured the models' performance with the AUC ROC metric (Bradley 1997). The metric was chosen given it is not sensitive to class imbalance and provides a threshold-independent estimate of models' discriminative capabilities. We did so in two different settings: binary classification and multiclass classification. While binary classification helps us understand how well the models discriminate whether there is a defect, the multiclass setting allows assessing how accurately the models learn to discriminate between specific types of defects.

We executed the experiments on two different machi nes: (a) a machine with four Intel Xeon Silver 4215R CPU processors with a 3.20GHz base frequency, with NVIDIA Tesla V100S-PCIE-32GB GPU and 31.4 GB of RAM; and (b) a machine with two Intel Xeon CPU processors with 2.3GHz base frequency, with a Tesla P100 16GB GPU and 13GB of RAM.

4.1. Experiment 1: few-shot learning on product images

The experiment compared how different models performed on the defect detection tasks. In particular, we were interested in comparing few-shot learning models and the classical supervised machine learning approach developed in Rožanec, Zajec, Theodoropoulos et al. (2022) and understanding the performance gap between both approaches. We also considered a SOTA few-shot learning model applied to defect detection and described in Takimoto et al. (2022). Furthermore, we were interested in how different pre-training and meta-training regimes influenced the few-shot models' performance. When training few-shot models, we used the PMF few-shot learning pipeline (Hu et al. 2022b), which achieved state-of-the-art results in various benchmarks. The experiment was performed on all the datasets listed in Section 3. Due to a lack of labelled samples, the experiments on the IBER dataset were only performed using one labelled image per class. We executed and compared the models described in Table 3.

4.2. Experiment 2: do DRAEM anomaly maps improve few-shot learning classifiers' performance?

Research by Rožanec, Zajec, Trajkova et al. (2022) has found that DRAEM anomaly maps boosted the performance of the classifiers. Therefore, we were interested in whether learning from anomaly maps could enhance the performance of few-shot learning models. To that end, we executed the same setup for Experiment 1 but considered two different inputs: (i) anomaly maps and (ii) the original product images with the corresponding anomaly maps. To combine images and anomaly maps into a single input, we concatenated feature vectors computed separately by the backbone model for the image and anomaly map. The experiment was run assuming a clean set of images of non-defective items existed to train a DRAEM

Table 3. Description of machine learning models used across the experiments.

Backbone	Pre train	Meta train	Type	Head
Resnet-50	cross-entropy	–	few-shot	ProtoNet
	DINO	–	few-shot	ProtoNet
ViT	cross-entropy	–	few-shot	ProtoNet
	DINO	–	few-shot	ProtoNet
		Meta-Dataset	few-shot	ProtoNet
	cross-entropy	MVTec-Capsule	few-shot	ProtoNet
Siamese	cross-entropy	MVTec-Capsule	few-shot	ProtoNet
Resnet18+MLP	cross-entropy	-	supervised	MLP

model and generate the anomaly maps. Given a large labelled set of such images contradicts the premises of few-shot learning, we devoted Experiment 3 to studying the effect of training DRAEM on noisy datasets.

4.3. Experiment 3: can we train DRAEM on impure datasets to generate anomaly maps without degrading the classifiers' performance?

The DRAEM model has been developed, assuming only images of non-defective products are provided. The model can, therefore, learn about what a non-defective product looks like and quickly identify whether some images are different from it and where the discrepancies exist. While in Experiment 2, we experiment with such a setting, the requirement to train a DRAEM model with only images of non-defective products contradicts one of the premises of the few-shot learning paradigm: only a few labelled examples exist for each class. Therefore, we were interested in whether the DRAEM model could be trained on all images and still produce some valuable output. In particular, we assumed the model could learn an average representation of the images and hint at any discrepancies in the anomaly map. While such discrepancies could no longer be identified with anomalies, they could still hint at how images from different classes differ between them, providing valuable information to determine their class. When training the DRAEM model, we considered three datasets (deco cap, shaft, and shavers, provided by *Philips Consumer Lifestyle BV*) and different degrees of imbalance (see Table 4). We trained the DRAEM models with default parameters until convergence. We performed multiclass classification with the ViT backbone model pre-trained with DINO and without meta-training to understand how the increasing impurity of the dataset on which the DRAEM models were trained affected the supervised classification model performance.

We trained few-shot classifiers considering the best model from Experiment 1 (ViT backbone pre-trained on ImageNet with DINO objective), the DRAEM anomaly maps as input, and trained with five samples per class.

4.4. Experiment 4: how can we construct a support set that maximises models' learning?

The selection of the support set in Experiment 1 was random. Few-shot learning aims to reduce the labelling effort by learning from only a few images shown to the model in training time. Therefore, this experiment aimed to understand how data selection can enhance the models' learning and the consequent classification results. To that end, we compared several well-known active learning strategies and developed novel active learning strategies too. In particular, we considered random sampling as a baseline data acquisition method. Among well-known approaches, we considered margin sampling (sample data instances where the difference between the top two most confident predictions is highest) and uncertainty sampling (sample data instances where the difference between the most confident prediction and absolute confidence is highest). We also developed three novel active learning techniques: (i) $\mathrm{EMB}_{\mathrm{ResNet-50}}$, (ii) $\mathrm{EMB}_{\mathrm{PMF}}$, and (iii) $\mathrm{EMB}_{\mathrm{GradCAM}}$. $\mathrm{EMB}_{\mathrm{ResNet-50}}$ computes the image embeddings using a ResNet-50 model pre-trained with a cross-entropy loss. Given a set of seed images of each class, it sources images furthest to them when computing the cosine distance. $\mathrm{EMB}_{\mathrm{PMF}}$ computes the image embeddings considering the PMF pipeline. The intuition behind this approach is that the PMF embeddings could be more discriminative than those obtained from a pre-trained ResNet model for a classification task. We considered a random set of seed images for each class to source unlabelled images. We looked for the unlabelled images that were furthest away, considering the cosine distance between embeddings. Finally, the $\mathrm{EMB}_{\mathrm{GradCAM}}$ technique followed a similar approach, considering the GradCAM heat map of the few-shot classification model trained on some random seed set. The GradCAM heat maps were computed for all the unlabelled images, and

Table 4. Dataset composition for different degrees of imbalance. The rate describes the number of defective samples we consider w.r.t. the original dataset.

	Class (count)									
DATASET	DECO CAP			SHAFT				SHAVERS		
Rate	Good	Flowlines	Marks	Good	Big	Small	Stripe	Good	Double	Interrupted
0.00	183	0	0	476	0	0	0	2409	0	0
0.01	183	2	2	476	24	9	2	2409	3	6
0.02	183	4	4	476	48	18	3	2409	5	11
0.05	183	9	9	476	118	43	7	2409	11	27
0.10	183	18	18	476	236	86	14	2409	22	54
0.20	183	36	35	476	471	172	28	2409	44	108
0.50	183	90	86	476	1178	430	68	2409	110	270

the cosine distance between their PMF embeddings was calculated, sourcing the images whose cosine distance was highest w.r.t. the seed image GradCAM heatmaps.

We trained few-shot classifiers considering the best model from Experiment 1 (ViT backbone pre-trained with DINO and no meta-training), with images as input, and trained them with one, five, ten, and twenty samples per class. We measured their performance in a multiclass setting.

5. Results

5.1. Experiment 1: few-shot learning outperformed a classical supervised machine learning model

We present the results of Experiment 1 in Tables 5 and 6 (binary classification) and Tables 7 and 8 (multiclass setting).

When performing binary classification with an image input, we observed that the best results were obtained in most cases with the ViT model pre-trained on ImageNet with DINO loss and without any meta-training. In particular, for one-shot learning, it displayed the best performance for the deco cap, shaft, and shavers datasets. The second-best performance for the shaft and shavers dataset was achieved with a ViT model pre-trained on ImageNet with DINO loss and meta-trained on the Meta-Dataset. On the other hand, the second-best performance was achieved for the deco cap dataset with a ViT model pre-trained on ImageNet with cross-entropy loss and meta-trained on MVTec-Capsule. This last model was the best when considering the IBER dataset. At the same time, the second-best performance was achieved with the Siamese backbone model pre-trained on ImageNet with cross-entropy loss and meta-trained on MVTec-Capsule. Five-shot learning increased the discriminative

Table 5. AUC ROC measured for classified images in a binary classification setting (defective vs. non-defective) for one-shot learning. The best results for each input type are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

	MODEL			$N = 1$			
INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers	IBER
IMAGE	Resnet50	ImageNet	–	0.6976 ± 0.0254	0.5741 ± 0.0145	0.6589 ± 0.0235	0.9998 ± 0.0009
		DINO	–	0.5376 ± 0.0157	0.5040 ± 0.0048	0.5194 ± 0.0092	0.7626 ± 0.0334
	ViT	ImageNet	–	0.6886 ± 0.0290	0.5102 ± 0.0925	0.5668 ± 0.0206	0.9995 ± 0.0004
		DINO	–	**1.0000 ± 0.0000**	**0.7009 ± 0.0582**	**0.7749 ± 0.0371**	0.9995 ± 0.0003
			Meta-Dataset	0.9943 ± 0.0106	*0.6339 ± 0.0796*	*0.7691 ± 0.0477*	0.9978 ± 0.0016
		Imagenet	MVTec-Capsule	*0.9999 ± 0.0001*	0.5656 ± 0.0210	0.7578 ± 0.0744	**0.9999 ± 0.0001**
	Siamese	ImageNet	MVTec-Capsule	0.7031 ± 0.0295	0.6204 ± 0.0186	0.6429 ± 0.0227	*0.9999 ± 0.0002*
	Resnet18 + MLP	ImageNet	–	0.9242 ± 0.0195	0.6082 ± 0.0403	0.6304 ± 0.0497	0.9950 ± 0.0034

Table 6. AUC ROC measured for classified images in a binary classification setting (defective vs. non-defective) for five-shot learning. The best results for each input type are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

	MODEL			$N = 5$		
INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers
IMAGE	Resnet50	ImageNet	–	0.8912 ± 0.0130	0.6657 ± 0.0125	*0.7868 ± 0.0131*
		DINO	–	0.5478 ± 0.0131	0.5450 ± 0.0167	0.5349 ± 0.0053
	ViT	ImageNet	–	0.8567 ± 0.0169	0.5147 ± 0.0091	0.6325 ± 0.0163
		DINO	–	*0.9960 ± 0.0074*	**0.7596 ± 0.0305**	**0.8075 ± 0.0340**
			Meta-Dataset	**0.9971 ± 0.0033**	0.7228 ± 0.0498	0.7859 ± 0.0382
		Imagenet	MVTec-Capsule	0.9986 ± 0.0023	0.5775 ± 0.0349	0.7815 ± 0.0400
	Siamese	ImageNet	MVTec-Capsule	0.8874 ± 0.0151	*0.7454 ± 0.0122*	0.7368 ± 0.0148
	Resnet18 + MLP	ImageNet	–	0.9803 ± 0.0045	0.7282 ± 0.0217	0.7317 ± 0.0587

Table 7. AUC ROC (one vs. rest) measured for classified images in a multiclass classification setting for one-shot learning. The best results for each input type are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

	MODEL			$N = 1$			
INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers	IBER
IMAGE	Resnet50	ImageNet	–	0.7942 ± 0.0193	0.5749 ± 0.0143	0.7068 ± 0.0181	0.9956 ± 0.0027
		DINO	–	0.5478 ± 0.0101	0.5049 ± 0.0021	0.5195 ± 0.0075	0.7594 ± 0.0331
	ViT	ImageNet	–	0.8002 ± 0.0222	0.5092 ± 0.0076	0.5911 ± 0.0196	0.9820 ± 0.0080
		DINO	–	**0.9757 ± 0.0113**	0.5788 ± 0.0486	**0.7317 ± 0.0287**	0.9989 ± 0.0005
			Meta-Dataset	0.9585 ± 0.0105	0.5660 ± 0.0465	*0.7308 ± 0.0565*	0.9977 ± 0.0017
		Imagenet	MVTec-Capsule	*0.9723 ± 0.0084*	0.5419 ± 0.0191	0.7241 ± 0.0608	**0.9994 ± 0.0003**
	Siamese	ImageNet	MVTec-Capsule	0.7956 ± 0.0221	**0.6244 ± 0.0168**	0.6986 ± 0.0167	*0.9990 ± 0.0005*
	Resnet18 + MLP	ImageNet	–	0.9587 ± 0.0116	*0.5925 ± 0.0213*	0.6441 ± 0.0465	0.9949 ± 0.0033

Table 8. AUC ROC (one vs. rest) measured for classified images in a multiclass classification setting for five-shot learning. The best results for each input type are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

	MODEL			$N = 5$		
INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers
IMAGE	Resnet50	ImageNet	–	0.9167 ± 0.0097	0.6470 ± 0.0132	**0.7994 ± 0.0106**
		DINO	–	0.5701 ± 0.0048	0.5300 ± 0.0092	0.5312 ± 0.0060
	ViT	ImageNet	–	0.9145 ± 0.0124	0.5075 ± 0.0076	0.6848 ± 0.0142
		DINO	–	**0.9947 ± 0.0011**	0.6646 ± 0.0280	*0.7936 ± 0.0199*
			Meta-Dataset	0.9871 ± 0.0031	0.6778 ± 0.0257	0.7538 ± 0.0371
		Imagenet	MVTec-Capsule	0.9865 ± 0.0032	0.5775 ± 0.0278	0.7268 ± 0.0248
	Siamese	ImageNet	MVTec-Capsule	0.9152 ± 0.0116	*0.7129 ± 0.0131*	0.7866 ± 0.0108
	Resnet18 + MLP	ImageNet	–	*0.9886 ± 0.0025*	**0.7183 ± 0.0158**	0.7373 ± 0.0547

power of the models. The ViT model pre-trained on ImageNet with DINO loss without any meta-training was the best classifier for the shaft and shavers datasets and the second-best for the deco cap dataset. For the model, we measured a perfect classification for the deco cap dataset in the one-shot learning setting, but we measured a slightly worse performance with five-shot learning. Among the second-best models, we found the ViT model pre-trained with DINO without any meta-training (deco cap dataset), the Siamese backbone model pre-trained on ImageNet with cross-entropy loss and meta-trained on MVTec-Capsule (shaft dataset), and the ResNet-50 backbone model pre-trained on ImageNet with cross-entropy loss without meta-training (shavers dataset). Comparing five-shot learning against one-shot learning, five-shot learning achieved a performance increase of 0.0587 and 0.0326 AUC ROC points at the shaft and shavers dataset when comparing the best models in each setting.

From the analysis above, we consider that the best performance was consistently delivered by the ViT backbone model pre-trained on ImageNet with DINO loss without meta-training. While the ResNet-18+MLP model always achieved competitive results, it never achieved the best performance and, in a few cases, was considered the second-best model for a given dataset. The Siamese network, pre-trained on ImageNet with cross-entropy loss and meta-trained on MVTec-Capsule, achieved the best and second-best performance in a few cases and remained competitive.

When switching to a multiclass setting, we observed the best and second-best performance was consistently delivered by the ViT backbone model pre-trained on ImageNet with cross-entropy loss without any meta-training. In particular, it achieved the best performance at the deco cap and shavers dataset when performing one-shot classification with images as the model's input. In five-shot learning, it achieved the best performance for the deco cap dataset (AUC ROC of 0.9947) and second-best for the shavers dataset (AUC ROC of 0.7936, increasing the discriminative performance by 0.0619 AUC ROC points w.r.t. the one-shot learning setting).

From the results and analysis above, we conclude that few-shot learning achieved better results than the model we compared to in a classical supervised machine learning setting. Furthermore, the ViT backbone model pre-trained on ImageNet with cross-entropy loss without any meta-training displayed a better performance than the Siamese network model described in Takimoto et al. (2022).

5.1.1. How are these results relevant to production systems?

Few-shot learning models achieved better performance than classical supervised machine learning models, confirming that good defect detection results can be achieved in a supervised setting using a small number of labelled instances. The fact that only a few labelled instances are required to train and test the model is of particular relevance, given it reduces the costs associated with searching and annotating such samples, reducing the time and effort required to start training a machine learning model for defect detection.

5.2. Experiment 2: DRAEM anomaly maps improve few-shot learning classifiers' performance

We present the results of Experiment 2 in Tables 9 and 10 (binary classification) and Tables 11 and 12 (multiclass setting).

In most cases, better results than those obtained when considering the image input data (Experiment 1) were achieved when training the models on DRAEM anomaly maps. In particular, for one-shot learning, the ViT model pre-trained on ImageNet with DINO loss without any meta-training was the best classifier for the shavers and IBER datasets. Nevertheless, four models achieved a perfect classification score for the IBER dataset. On the other hand, the ViT backbone model pre-trained on ImageNet with DINO loss and meta-trained on the Meta-Dataset was considered best for the deco cap and shaft datasets. The second-best classification model for the deco cap and shaft datasets was the Resnet18+MLP pre-trained

Table 9. AUC ROC measured for models in a binary classification setting (defective vs. non-defective) for one-shot learning. The best results are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

	MODEL			$N = 1$			
INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers	IBER
ANOMALY MAP	Resnet50	Imagenet	–	0.9543 ± 0.0195	0.5930 ± 0.0169	*0.8091 ± 0.0272*	0.9884 ± 0.0056
		DINO	–	0.6404 ± 0.0089	0.5056 ± 0.0025	0.5500 ± 0.0050	0.7827 ± 0.0268
	ViT	Imagenet	–	0.9894 ± 0.0047	0.5520 ± 0.0136	0.6974 ± 0.0248	**1.0000 ± 0.0000**
		DINO	–	0.9995 ± 0.0008	0.6328 ± 0.0467	**0.8261 ± 0.0613**	**1.0000 ± 0.0000**
			Meta-Dataset	**0.9999 ± 0.0001**	**0.6707 ± 0.0709**	0.7549 ± 0.0827	**1.0000 ± 0.0000**
		ImageNet	MVTec-Capsule	0.9884 ± 0.0075	0.6395 ± 0.0754	0.7624 ± 0.0753	*0.9991 ± 0.0009*
	Siamese	ImageNet	MVTec-Capsule	0.9559 ± 0.0131	0.5603 ± 0.0158	0.6392 ± 0.0191	**1.0000 ± 0.0000**
	Resnet18 + MLP	ImageNet	–	*0.9999 ± 0.0002*	*0.6424 ± 0.0473*	0.7712 ± 0.1011	0.9989 ± 0.0010
IMAGE + ANOMALY MAP	Resnet50	Imagenet	–	0.6195 ± 0.0189	0.5303 ± 0.0097	0.5984 ± 0.0174	0.6662 ± 0.0192
		DINO	–	0.8623 ± 0.0420	0.5119 ± 0.0079	0.5908 ± 0.0148	0.9732 ± 0.0096
	ViT	Imagenet	–	0.5686 ± 0.0157	0.5080 ± 0.0072	0.5590 ± 0.0137	0.6993 ± 0.0180
		DINO	–	*0.9998 ± 0.0003*	**0.6916 ± 0.0645**	**0.8417 ± 0.0569**	**1.0000 ± 0.0000**
			Meta-Dataset	**0.9999 ± 0.0002**	*0.6888 ± 0.0623*	*0.8033 ± 0.0173*	**1.0000 ± 0.0000**
		ImageNet	MVTec-Capsule	0.9924 ± 0.0064	0.6382 ± 0.0603	0.7756 ± 0.0699	**1.0000 ± 0.0000**
	Siamese	ImageNet	MVTec-Capsule	0.6488 ± 0.0182	0.5494 ± 0.0107	0.5477 ± 0.0096	0.6441 ± 0.0211
	Resnet18 + MLP	ImageNet	–	0.9994 ± 0.0011	0.6628 ± 0.0339	0.7544 ± 0.0979	*0.9999 ± 0.0001*

Table 10. AUC ROC measured for models in a binary classification setting (defective vs. non-defective) for five-shot learning. The best results are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

	MODEL			$N = 5$		
INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers
ANOMALY MAP	Resnet50	Imagenet	–	0.9918 ± 0.0017	0.6925 ± 0.0131	*0.9096 ± 0.0087*
		DINO	–a	0.6657 ± 0.0101	0.5132 ± 0.0029	0.5616 ± 0.0053
	ViT	Imagenet	–	*0.9996 ± 0.0004*	0.6207 ± 0.0128	0.8204 ± 0.0118
		DINO	–	**1.0000 ± 0.0000**	0.7232 ± 0.0277	**0.9452 ± 0.0238**
			Meta-Dataset	**1.0000 ± 0.0000**	*0.7661 ± 0.0451*	0.8851 ± 0.0238
		ImageNet	MVTec-Capsule	0.9966 ± 0.0025	**0.7771 ± 0.0209**	0.9019 ± 0.0457
	Siamese	ImageNet	MVTec-Capsule	0.9944 ± 0.0015	0.6341 ± 0.0132	0.7388 ± 0.0180
	Resnet18 + MLP	ImageNet	–	**1.0000 ± 0.0000**	0.7459 ± 0.0202	0.8947 ± 0.0301
IMAGE + ANOMALY MAP	Resnet50	Imagenet	–	0.6393 ± 0.0156	0.5519 ± 0.0093	0.6575 ± 0.0145
		DINO	–	0.9149 ± 0.0146	0.5508 ± 0.0209	0.6442 ± 0.0086
	ViT	Imagenet	–	0.5888 ± 0.0187	0.5158 ± 0.0065	0.6107 ± 0.0122
		DINO	–	**1.0000 ± 0.0000**	0.7871 ± 0.0245	**0.9568 ± 0.0200**
			Meta-Dataset	**1.0000 ± 0.0000**	**0.7975 ± 0.0228**	0.8974 ± 0.0242
		ImageNet	MVTec-Capsule	0.9992 ± 0.0006	0.7419 ± 0.0289	*0.9098 ± 0.0412*
	Siamese	ImageNet	MVTec-Capsule	0.6977 ± 0.0179	0.5653 ± 0.0094	0.5597 ± 0.0089
	Resnet18 + MLP	ImageNet	–	*0.9999 ± 0.0001*	*0.7912 ± 0.0202*	0.8960 ± 0.0275

on ImageNet with cross-entropy loss, which achieved almost the same performance as the best model for the deco cap dataset and lagged less than 0.03 AUC ROC points behind the best classifier for the shaft dataset. The second-best models for the shavers and IBER dataset were the ResNet-50 backbone model pre-trained on ImageNet with cross-entropy loss and the ViT backbone model pre-trained on ImageNet with cross-entropy loss and meta-trained on MVTec-Capsule, respectively. Five-shot learning increased the models' discriminative performance. In particular, the ViT backbone model pre-trained on ImageNet with DINO loss and without any meta-training was considered the best classifier for the deco cap and shavers datasets. A perfect classification performance on the deco cap dataset was achieved by two additional models: the ViT backbone model pre-trained on ImageNet with DINO loss and meta-trained on Meta-Dataset; and the ResNet-18+MLP model. The ViT backbone model achieved the second-best performance pre-trained on the ImageNet dataset with cross-entropy loss without meta-training. On the other hand, the second-best model at the shaft dataset was the ViT backbone model pre-trained on ImageNet with DINO loss and without any meta-training. The first and second-best models at the shavers dataset remained the same as for one-shot learning. Comparing five-shot learning against one-shot learning, five-shot learning achieved a performance increase of 0.1064 and 0.1191 AUC ROC points at the shaft and shavers dataset when comparing the best models in each setting.

Consistent with the findings of Rožanec, Zajec, Theodoropoulos et al. (2022), the best results were achieved when considering the image and anomaly map as inputs to the machine learning model. In particular, the best classifier for the shaft, shavers, and IBER datasets for one-shot learning was the ViT model pre-trained on

Table 11. AUC ROC (one-vs-rest) measured for models in a multiclass classification setting for one-shot learning. The best results are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

	MODEL			$N = 1$			
INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers	IBER
ANOMALY MAP	Resnet50	Imagenet	–	0.9749 ± 0.0131	**0.5998 ± 0.0161**	**0.8427 ± 0.0243**	0.9907 ± 0.0062
		DINO	–	0.6564 ± 0.0083	0.5041 ± 0.0017	0.5436 ± 0.0041	0.7804 ± 0.0271
	ViT	Imagenet	–	0.9301 ± 0.0120	0.5549 ± 0.0119	0.7054 ± 0.0226	0.9980 ± 0.0008
		DINO	–	*0.9938 ± 0.0030*	0.5364 ± 0.0374	*0.8087 ± 0.0600*	**0.9999 ± 0.0001**
			Meta-Dataset	**0.9939 ± 0.0025**	0.5552 ± 0.0320	0.7467 ± 0.0778	*0.9997 ± 0.0002*
		ImageNet	MVTec-Capsule	0.9764 ± 0.0082	0.5495 ± 0.0255	0.7529 ± 0.0733	0.9988 ± 0.0010
	Siamese	ImageNet	MVTec-Capsule	0.9686 ± 0.0150	*0.5811 ± 0.0159*	0.6975 ± 0.0196	0.9908 ± 0.0084
	Resnet18 + MLP	ImageNet	–	0.9931 ± 0.0037	0.5669 ± 0.0191	0.7651 ± 0.0939	0.9989 ± 0.0010
IMAGE + ANOMALY MAP	Resnet50	Imagenet	–	0.6430 ± 0.0193	0.5333 ± 0.0072	0.6006 ± 0.0159	0.7888 ± 0.0223
		DINO	–	0.8694 ± 0.0303	0.5129 ± 0.0048	0.5849 ± 0.0124	0.9710 ± 0.0094
	ViT	Imagenet	–	0.5639 ± 0.0195	0.5208 ± 0.0054	0.5592 ± 0.0135	0.8181 ± 0.0222
		DINO	–	*0.9962 ± 0.0024*	0.5711 ± 0.0475	**0.8219 ± 0.0540**	**1.0000 ± 0.0000**
			Meta-Dataset	**0.9970 ± 0.0019**	**0.5776 ± 0.0413**	*0.7637 ± 0.0715*	**1.0000 ± 0.0000**
		ImageNet	MVTec-Capsule	0.9838 ± 0.0072	0.5559 ± 0.0233	0.7628 ± 0.0677	0.9991 ± 0.0006
	Siamese	ImageNet	MVTec-Capsule	0.6469 ± 0.0208	0.5441 ± 0.0082	0.5497 ± 0.0093	0.7356 ± 0.0263
	Resnet18 + MLP	ImageNet	–	0.9944 ± 0.0034	*0.5751 ± 0.0164*	0.7580 ± 0.0913	*0.9998 ± 0.0001*

Table 12. AUC ROC (one-vs-rest) measured for models in a multiclass classification setting for five-shot learning. The best results are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

	MODEL			$N = 5$		
INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers
ANOMALY MAP	Resnet50	Imagenet	–	0.9923 ± 0.0001	**0.6988 ± 0.0117**	**0.9381 ± 0.0070**
		DINO	–	0.6815 ± 0.0062	0.5070 ± 0.0023	0.5513 ± 0.0046
	ViT	Imagenet	–	0.9894 ± 0.0024	0.6044 ± 0.0110	0.8698 ± 0.0091
		DINO	–	0.9941 ± 0.0027	0.6013 ± 0.0122	*0.9243 ± 0.0259*
			Meta-Dataset	*0.9954 ± 0.0023*	0.6310 ± 0.0155	0.8710 ± 0.0210
		ImageNet	MVTec-Capsule	0.9810 ± 0.0073	0.6299 ± 0.0165	0.8907 ± 0.0438
	Siamese	ImageNet	MVTec-Capsule	**0.9958 ± 0.0012**	*0.6767 ± 0.0126*	0.8038 ± 0.0161
	Resnet18 + MLP	ImageNet	–	0.9941 ± 0.0028	0.6601 ± 0.0295	0.8821 ± 0.0295
IMAGE + ANOMALY MAP	Resnet50	Imagenet	–	0.6522 ± 0.0184	0.5649 ± 0.0070	0.6688 ± 0.0150
		DINO	–	0.8969 ± 0.0132	0.5454 ± 0.0129	0.6341 ± 0.0084
	ViT	Imagenet	–	0.5806 ± 0.0183	0.5331 ± 0.0057	0.6349 ± 0.0126
		DINO	–	*0.9962 ± 0.0020*	*0.6658 ± 0.0214*	**0.9337 ± 0.0224**
			Meta-Dataset	**0.9981 ± 0.0012**	**0.6977 ± 0.0178**	0.8844 ± 0.0206
		ImageNet	MVTec-Capsule	0.9868 ± 0.0054	0.6350 ± 0.0218	*0.8962 ± 0.0403*
	Siamese	ImageNet	MVTec-Capsule	0.6600 ± 0.0173	0.5655 ± 0.0076	0.5812 ± 0.0095
	Resnet18 + MLP	ImageNet	–	0.9950 ± 0.0026	0.6501 ± 0.0191	0.8924 ± 0.0266

ImageNet with DINO loss without any meta-training. Furthermore, this model was the second-best in the deco cap dataset. On the other hand, the ViT model pre-trained on ImageNet with DINO loss and meta-trained on the Meta-Dataset achieved the best performance on the deco cap and IBER datasets and the second-best performance on the shaft and shavers datasets. Five-shot learning also led to better results in this case. The ViT model pre-trained on ImageNet with DINO loss without any meta-training achieved the best performance on the deco cap and shavers datasets, while the ViT model pre-trained on ImageNet with DINO loss and meta-trained on the Meta-Dataset achieved the best performance on the deco cap and shaft datasets. Comparing five-shot learning against one-shot learning, five-shot learning achieved a performance increase of 0.1059 and 0.1151 AUC ROC points at the shaft and shavers dataset when comparing the best models in each setting.

In the multiclass one-shot setting, when the models' input consisted of anomaly maps, the ViT backbone model pre-trained on ImageNet with cross-entropy loss without any meta-training, achieved the best performance among models for the IBER dataset and second-best among the models developed for the deco cap and shavers datasets. The ResNet-50 model pre-trained on ImageNet with cross-entropy loss achieved the best performance among models for the shaft and shavers dataset. This remained true for the five-shot learning models. The ViT backbone model pre-trained on ImageNet with cross-entropy loss and meta-trained on Meta-Dataset was best for the deco cap dataset in the one-shot learning setting. Nevertheless, in a five-shot learning setting, the Siamese model pre-trained on ImageNet with cross-entropy loss and meta-trained on MVTec-Capsule issued the best results while securing its position as the second-best model for the shaft dataset.

Considering image and anomaly map inputs to the classification model, we noticed that the discriminative power in some cases increased w.r.t. the models whose input was only an anomaly map. Nevertheless, this was not always the case, as observed in the binary classification case. The ViT backbone model pre-trained on ImageNet with cross-entropy loss and without meta-training was the best or second-best model in most cases, either for one-shot or five-shot settings. The only exception was the shaft dataset in the one-shot setting, where the ViT backbone model achieved the best performance pre-trained on ImageNet with cross-entropy loss and meta-trained on Meta-Dataset, while the second-best performance was attributed to the ResNet-18+MLP model. The ViT backbone model pre-trained on ImageNet with cross-entropy loss and meta-trained on Meta-Dataset achieved the best and second-best performance in all but one case (the shavers dataset in the five-shot learning setting).

In general, it is observed that training the model in a five-shot setting achieves better results than in a one-shot setting. E.g. the ResNet-50 backbone model trained on ImageNet with cross-entropy loss without meta-training achieved the best performance in one and five-shot learning when trained with anomaly maps. Changing from a one-shot to a five-shot setting increased its performance by 0.099 AUC ROC points when predicting the shaft dataset and 0.0954 AUC ROC points for the shavers dataset. Similarly, the ViT backbone model pre-trained on ImageNet with cross-entropy loss without meta-training increased its performance by 0.1118 AUC ROC points for the shavers dataset when trained with images and anomaly maps. On the other hand, the ViT backbone model pre-trained on ImageNet with cross-entropy loss and meta-trained on Meta-Dataset increased its performance by 0.1201 AUC ROC points in the shaft dataset and went to achieve perfect classification for the deco cap dataset.

When comparing the results obtained for best models across Experiment 1 and Experiment 2 for the multi-class setting, the models trained with anomaly maps or anomaly maps and images achieved superior results with two exceptions: the shaft dataset with one and five-shot learning.

From the results obtained in Experiment 1 and Experiment 2, we conclude that few-shot learning provides the best results when leveraging anomaly maps. Anomaly maps provide richer information to the classifier by highlighting where potential defects exist, easing the learning process. Using images and anomaly maps or only anomaly maps depends on the classification setting. Experiment 1 and Experiment 2 show that it always issued better results using anomaly maps and images for binary classification, which was not always true for multiclass settings. The best performance was consistently delivered by the ViT backbone model pre-trained on ImageNet with DINO without meta-training, regardless of the input features used to classify the images. While the ResNet-18+MLP model always achieved competitive results, it never reached the best performance and, in a few cases, was considered the second-best model for a particular dataset. The Siamese network, pre-trained on ImageNet with cross-entropy loss and meta-trained on MVTec-Capsule, achieved the best and second-best performance in a few cases but significantly lagged behind the best models in many others, lacking a consistent performance across datasets and experimental settings.

5.2.1. How are these results relevant to production systems?

The anomaly maps should display regions where the potential defects could be located. Such representation seems to favorably affect the learning of machine learning models, which achieved better outcomes leveraging anomaly maps or images and anomaly for defect detection. Therefore, when developing machine learning models for defect detection, machine learning engineers should consider how such anomaly maps could be created and leveraged to enhance the defect detection outcomes against those that could be achieved by leveraging the product images only.

5.3. Experiment 3: training DRAEM anomaly maps on impure datasets can affect the classifier's performance

We summarise the results of this experiment in Table 13. When training the DRAEM model with different impurity levels, we observed that increasing the impurity level degraded the few-shot learning model's discriminative performance. While for the deco cap and shaft datasets, the performance achieved by the few-shot learning model on anomaly maps was lower than that achieved when trained on images, this was not the case for the shavers dataset. In this last case, we observed that training a DRAEM model on an increasing impurity level still provided an advantage over the models trained directly on the images. In particular, a few-shot learning model trained on DRAEM anomaly maps outperformed those trained on images, even when the dataset used to train the DRAEM model had 10% of defective samples. While the results do not provide conclusive evidence, further research is required to understand what image characteristics enable using impure datasets when training DRAEM models and still benefit from the DRAEM insights to outperform models trained solely on images.

Table 13. AUC ROC measured for multiclass classification, considering DRAEM anomaly maps as inputs. The anomaly maps are obtained from a DRAEM model trained with all available images. We specify different degrees of artificial imbalance on top of the given one (mix rate). The exact composition of the datasets for each mix rate is described in Table 4. We report the 95% confidence intervals. The models using anomaly map features are bolded when performing better than those trained solely on images.

	MODEL						
INPUT FEATURES	Backbone	Pre train	Meta train	Mix rate	Deco cap	Shaft	Shavers
IMAGE	ViT	DINO	–	**1.00**	0.9947 ± 0.0011	0.6646 ± 0.0280	0.7936 ± 0.0199
ANOMALY MAP	ViT	DINO	–	**0.00**	0.9941 ± 0.0027	0.6013 ± 0.0122	**0.9243 ± 0.0259**
				0.01	0.9811 ± 0.0079	0.5726 ± 0.0242	**0.9213 ± 0.0211**
				0.02	0.9408 ± 0.0122	0.5487 ± 0.0187	**0.9111 ± 0.0250**
				0.05	0.8815 ± 0.0062	0.5213 ± 0.0019	**0.9057 ± 0.0159**
				0.10	0.7515 ± 0.0293	0.5018 ± 0.0007	**0.8570 ± 0.0159**
				0.20	0.6004 ± 0.0018	0.5010 ± 0.0015	0.6413 ± 0.0211
				0.50	0.5102 ± 0.0023	0.5000 ± 0.0001	0.5020 ± 0.0391

5.3.1. How are these results relevant to production systems?

The experiment does not provide conclusive results on when using impure datasets can affect the quality of the DRAEM model outcomes. Nevertheless, when impure datasets allow for good-quality outcomes, little or no manual annotation is required to create such datasets, preserving the benefits of the few-shot learning while enabling superior results compared to training the few-shot learning models only with images.

5.4. Experiment 4: active learning techniques provide effective means to construct support sets that maximise the models' learning

We compared six active learning techniques for this experiment and applied them only to select the images on which the few-shot models were trained. We present the results in Table 14 and Figure 4. The margin sampling technique achieved the best performance for the deco cap dataset in most cases. The only exception was when twenty samples were shown to the few-shot learning algorithm, where uncertainty sampling was best. Uncertainty sampling was also second-best in the rest of the cases. In most cases, the custom active learning techniques beat at least one widely adopted sampling method (random, uncertain, or margin sampling) but never achieved the best or second-best performance. Uncertainty sampling also performed best when selecting five or ten samples from the shavers dataset and was second-best when selecting twenty. Nevertheless, margin sampling was best when selecting twenty samples, and random sampling outperformed other methods when selecting a single sample while being the second-best when considering five or ten data samples. It is worth mentioning that the EMB$_{\text{GradCAM}}$ technique achieved the second-best performance if selecting a single sample. The shaft dataset displayed different dynamics. The best sampling method for one and five samples was uncertainty sampling, followed by random sampling (when selecting

Table 14. AUC ROC measured for models in a multiclass classification setting. The best results are bolded, and the second-best results are highlighted in italics. We report the 95% confidence intervals.

		Number of sampled images			
Dataset	Active learning strategy	1	5	10	20
DECO CAP	Random sampling	0.9437 ± 0.0252	0.9643 ± 0.0135	0.9811 ± 0.0052	*0.9912 ± 0.0029*
	Uncertainty sampling	*0.9506 ± 0.0226*	*0.9880 ± 0.0026*	*0.9959 ± 0.0011*	**0.9994 ± 0.0002**
	Margin sampling	**0.9533 ± 0.0230**	**0.9889 ± 0.0031**	**0.9969 ± 0.0008**	0.9708 ± 0.0375
	EMB$_{\text{ResNet-50}}$	0.9467 ± 0.0301	0.9690 ± 0.0067	0.9735 ± 0.0054	0.9881 ± 0.0027
	EMB$_{\text{PMF}}$	0.9485 ± 0.0300	0.9762 ± 0.0052	0.9761 ± 0.0050	0.9786 ± 0.0042
	EMB$_{\text{GradCAM}}$	0.9455 ± 0.0469	0.9790 ± 0.0088	0.9883 ± 0.0052	0.9934 ± 0.0032
SHAFT	Random sampling	*0.6210 ± 0.0667*	0.6427 ± 0.0686	**0.7320 ± 0.0637**	0.7612 ± 0.0212
	Uncertainty sampling	**0.6266 ± 0.0686**	**0.6978 ± 0.0326**	0.7038 ± 0.0164	0.7624 ± 0.0239
	Margin sampling	0.6023 ± 0.0747	0.6833 ± 0.0216	0.6838 ± 0.0518	0.7406 ± 0.0363
	EMB$_{\text{ResNet-50}}$	0.5965 ± 0.0663	0.6435 ± 0.0442	0.6889 ± 0.0170	**0.7870 ± 0.0194**
	EMB$_{\text{PMF}}$	0.6099 ± 0.0333	*0.6866 ± 0.0444*	*0.7154 ± 0.0664*	*0.7667 ± 0.0492*
	EMB$_{\text{GradCAM}}$	0.6075 ± 0.0525	0.6573 ± 0.0354	0.7014 ± 0.0326	0.7519 ± 0.0239
SHAVERS	Random sampling	**0.6987 ± 0.0312**	*0.7206 ± 0.0289*	*0.7582 ± 0.0299*	0.7934 ± 0.0251
	Uncertainty sampling	0.6541 ± 0.0303	**0.7271 ± 0.0236**	**0.7679 ± 0.0187**	*0.7979 ± 0.0169*
	Margin sampling	0.6687 ± 0.0256	0.7226 ± 0.0346	0.7519 ± 0.0293	**0.8129 ± 0.0199**
	EMB$_{\text{ResNet-50}}$	0.6601 ± 0.0332	0.6621 ± 0.0312	0.6865 ± 0.0252	0.7205 ± 0.0385
	EMB$_{\text{PMF}}$	0.6664 ± 0.0323	0.7072 ± 0.0218	0.7371 ± 0.0199	0.7813 ± 0.0237
	EMB$_{\text{GradCAM}}$	*0.6748 ± 0.0390*	0.7078 ± 0.0447	0.7023 ± 0.0375	0.7508 ± 0.0556

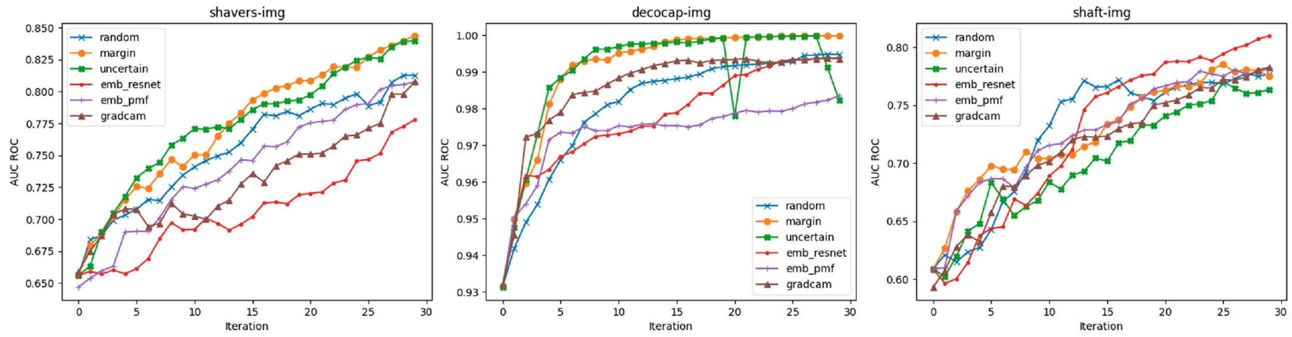

Figure 4. Plots showing AUC ROC performance based on sampling iterations for six active learning techniques and three datasets. The following naming conventions are used for the series: random (random sampling), margin (margin sampling), uncertain (uncertainty sampling), emb_resnet ($EMB_{ResNet-50}$), emb_pmf (EMB_{PMF}), and GradCAM ($EMB_{GradCAM}$).

a single sample) and EMB_{PMF} when selecting five samples. EMB_{PMF} was also second-best when selecting ten and twenty samples, second only to random sampling (when selecting ten samples) and $EMB_{ResNet-50}$ (when sampling twenty data samples).

The results presented in Table 14 and the analysis provided above indicate that following a particular sampling technique rather than a random sampling of the data can be beneficial for the model's learning process and lead to a higher discriminative power for the same amount of data instances shown to the few-shot learning model.

5.4.1. How are these results relevant to production systems?

Active learning techniques provide means to select unlabelled data that could lead to better learning of machine learning models. Using such techniques reduces the data labelling cost, and a higher machine-learning model performance is achieved. The proposed novel active learning techniques show a promising performance and could help reduce the manual annotation effort in the context of developing a machine learning defect detection solution.

6. Discussion

One of this research's main innovations is using heatmaps generated with unsupervised machine learning models to enhance the performance of supervised models. In particular, it was explored how few-shot learning classification models can be enhanced by enriching the input (product images) with anomaly maps obtained for such products from unsupervised machine learning models. We consider the results supporting the abovementioned hypothesis (see Tables 15 and 16) as one of the main contributions of this research. The abovementioned tables show that in most cases, the models' performance is enhanced at one-shot learning, but certainly in five-shot learning scenarios.

Furthermore, we studied whether the anomaly maps built with the DRAEM method require good samples to train the model or whether a similar performance could be achieved in the few-shot learning setting if such anomaly maps are trained on impure datasets, regardless of the class to which the data instances belong. Our results show that impure datasets may lead to a degraded classifier performance. Nevertheless, the impurity level affecting the DRAEM anomaly maps and the

Table 15. Comparison of best models, considering Image (I) and Image+Heatmap (I+H) input features. The results were taken from Table 5 and Table 6 regarding the binary classification case.

		MODEL			DATASET			
N	INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers	IBER
1	I	ViT	DINO	–	**1.0000 ± 0.0000**	**0.7009 ± 0.0582**	0.7749 ± 0.0371	0.9995 ± 0.0003
				Meta-Dataset	0.9943 ± 0.0106	0.6339 ± 0.0796	0.7691 ± 0.0477	0.9978 ± 0.0016
			Imagenet	MVTec-Capsule	**0.9999 ± 0.0001**	0.5656 ± 0.0210	0.7578 ± 0.0744	0.9999 ± 0.0001
	I+H	ViT	DINO	–	0.9998 ± 0.0003	0.6916 ± 0.0645	**0.8417 ± 0.0569**	**1.0000** ± 0.0000
				Meta-Dataset	**0.9999 ± 0.0002**	**0.6888 ± 0.0623**	**0.8033 ± 0.0173**	**1.0000 ± 0.0000**
			ImageNet	MVTec-Capsule	0.9924 ± 0.0064	**0.6382 ± 0.0603**	**0.7756 ± 0.0699**	**1.0000 ± 0.0000**
5	I	ViT	DINO	–	0.9960 ± 0.0074	0.7596 ± 0.0305	0.8075 ± 0.0340	–
				Meta-Dataset	0.9971 ± 0.0033	0.7228 ± 0.0498	0.7859 ± 0.0382	–
	I+H	ViT	DINO	–	**1.0000 ± 0.0000**	**0.7871 ± 0.0245**	**0.9568 ± 0.0200**	–
				Meta-Dataset	**1.0000 ± 0.0000**	**0.7975 ± 0.0228**	**0.8974 ± 0.0242**	–

Table 16. Comparison of best models, considering Image (I) and Image+Heatmap (I+H) input features. The results were taken from Table 7 and Table 8 regarding the multiclass classification case.

		MODEL			DATASET			
N	INPUT FEATURES	Backbone	Pre train	Meta train	Deco cap	Shaft	Shavers	IBER
1	I	ViT	DINO	–	0.9757 ± 0.0113	**0.5788 ± 0.0486**	0.7317 ± 0.0287	0.9989 ± 0.0005
				Meta-Dataset	0.9585 ± 0.0105	0.5660 ± 0.0465	0.7308 ± 0.0565	0.9977 ± 0.0017
			Imagenet	MVTec-Capsule	0.9723 ± 0.0084	0.5419 ± 0.0191	0.7241 ± 0.0608	**0.9994 ± 0.0003**
		Siamese	ImageNet	MVTec-Capsule	**0.7956 ± 0.0221**	**0.6244 ± 0.0168**	**0.6986 ± 0.0167**	**0.9990 ± 0.0005**
	I+H	ViT	DINO	–	**0.9962 ± 0.0024**	0.5711 ± 0.0475	**0.8219 ± 0.0540**	**1.0000 ± 0.0000**
				Meta-Dataset	**0.9970 ± 0.0019**	**0.5776 ± 0.0413**	**0.7637 ± 0.0715**	**1.0000 ± 0.0000**
			ImageNet	MVTec-Capsule	**0.9838 ± 0.0072**	**0.5559 ± 0.0233**	**0.7628 ± 0.0677**	0.9991 ± 0.0006
		Siamese	ImageNet	MVTec-Capsule	0.6469 ± 0.0208	0.5441 ± 0.0082	0.5497 ± 0.0093	0.7356 ± 0.0263
5	I	ViT	DINO	–	0.9947 ± 0.0011	0.6646 ± 0.0280	0.7936 ± 0.0199	–
				Meta-Dataset	0.9871 ± 0.0031	0.6778 ± 0.0257	0.7538 ± 0.0371	–
		Resnet18 + MLP	ImageNet	–	0.9886 ± 0.0025	**0.7183 ± 0.0158**	0.7373 ± 0.0547	–
	I+H	ViT	DINO	–	**0.9962 ± 0.0020**	**0.6658 ± 0.0214**	**0.9337 ± 0.0224**	–
				Meta-Dataset	**0.9981 ± 0.0012**	**0.6977 ± 0.0178**	**0.8844 ± 0.0206**	–
		Resnet18 + MLP	ImageNet	–	**0.9950 ± 0.0026**	0.6501 ± 0.0191	**0.8924 ± 0.0266**	–

corresponding classifier can vary among datasets. This remains an open question, and further experiments are required to determine whether similar shortcomings and behaviours are observed across a broader range of unsupervised defect detection methods. While this could be considered a drawback to a few-shot learning approach, given it may require labelling large amounts of images for non-defective items, the impact may not be critical: most products correspond to products without a defect. This could facilitate the acquisition of a dataset of images for defect-free items.

Finally, we have shown that active learning techniques in few-shot learning can boost models' learning and lead to better outcomes. Three novel active learning methods were proposed: EMB$_{\text{ResNet}-50}$, EMB$_{\text{PMF}}$, and EMB$_{\text{GradCAM}}$. The three methods showed promising results, beating widely adopted active learning sampling methods. Nevertheless, in all but one case, the proposed methods did not lead to the best performance of the few-shot learning models. Further research is required to understand how these methods can be evolved and enhanced. Furthermore, little research has been performed on the intersection of explainable artificial intelligence and active learning. E.g. Ghai et al. (2021) leveraged local explanations to help annotators annotate unlabelled data instances, and Ciravegna et al. (2023) studied how rule-based (domain or explainable artificial intelligence) knowledge can be converted into logic constraints and their violation checked to guide sample selection. Another active learning approach leveraging explainable artificial intelligence on images has been developed by Križnar et al. (2023). Therefore, the EMB$_{\text{GradCAM}}$ joins the reduced number of active learning methods for images that leverage insights from explainable artificial intelligence to guide the data sampling process. Having performed our research on four different datasets and observed a consistent behaviour of our models, we are confident our findings can be extrapolated to other visual inspection settings.

7. Conclusions and future work

This research explored using few-shot learning for automated visual inspection across four real-world datasets. The results show that few-shot learning models outperformed a regular machine learning classifier. Furthermore, few-shot learning models, whose input is images and anomaly maps, achieve stronger discriminative performance than few-shot learning models trained only on images for defect classification. Nevertheless, such anomaly maps may require annotating a dataset of good samples, defeating the purpose of few-shot learning. Our results show that training DRAEM unsupervised classification models on impure datasets to generate anomaly maps without prior data annotation does not guarantee informative ones. On the other hand, active learning strategies could be used to avoid random sampling data and instead obtain an annotated dataset that maximises the models' learning and discriminative performance. The results also confirmed that the models' discriminative capability greatly improved when considering five-shot learning against one-shot learning. While the classification models could discriminate whether a manufactured piece was defective, the performance usually decreased when determining the defect type. We consider the results to be promising. Nevertheless, in some cases, further effort is required to enhance them and ensure they satisfy manufacturing quality acceptance levels. Future work will focus on (i) new sampling techniques that allow for better reuse of the few labelled samples across episodes in the training set, (ii) new sampling techniques that consider characteristics of images and anomaly maps to create an initial support set in few-shot learning settings, and (iii) using few-shot generative adversarial networks to increase the amount of data in the support set (which we expect would enhance the models' performance with no additional labelling effort).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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

Notes on contributors

Patrik Zajec is a Ph.D. candidate at the Jožef Stefan International Postgraduate School. His research is mainly in the field of natural language processing, more specifically, he develops methods for unsupervised and semi-supervised information extraction and retrieval. He is also involved in the European Union Horizon 2020 project STAR, where he works on synthetic data augmentation techniques and a human-machine collaboration system using active learning, XAI, and semantic technologies.

Jože M. Rožanec Ph.D. is a post-doc researcher at the Artificial Intelligence Laboratory (Jožef Stefan Institute), and a machine learning engineer at Qlector d.o.o. (developing intelligent solutions for smart factories). He collaborates with the American Slovenian Education Foundation, where he leads multiple activities for Fellows and Alumni. Over more than ten years, he worked for several companies (e.g. Mercado Libre, Navent, Globant) in software engineering and machine learning-related roles. His research interests include machine learning methods for recommendations, fraud detection, demand forecasting, active learning, and explainable artificial intelligence (XAI).

Spyros Theodoropoulos is a Ph.D. Candidate at the School of Electrical and Computer Engineering of the National Technical University of Athens. He graduated from the same school and held an MSc in Machine Learning from Imperial College London. He has worked in the industry as a Software and Big Data Engineer and is now a member of the Data & Cloud Research Group at the University of Piraeus. His research is focused on deep learning, reinforcement learning, and the use of simulation for their deployment in real-life dynamic environments.

Mihail Fontul has a degree in mechanical engineering from the Faculty of Mechanics of Technical University of Cluj-Napoca, Romania, PhD in mechanical engineering from Instituto Superior Técnico of the Technical University of Lisbon, Portugal, and has a vast and balanced industrial and academic experience. He served as Assistant Professor at the Department of Mechanical Engineering at Instituto Superior Técnico, where taught and developed scientific activities in the areas of Mechanical Vibrations and Noise, Product Development, Design and Materials in Engineering. He is the author or co-author of several articles in journals and conferences and has also been awarded two times for academic achievement. He is currently responsible for the Research, Development and Innovation Department of the company IBER-OLEFF Componentes Técnicos em Plástico SA and managing partner of the company FONTUL LDA dedicated to technical-scientific engineering consultancy activities.

Erik Koehorst (Ir.) is a project manager currently working for Philips. He has studied industrial engineering management at the university of Twente and mechanical engineering at HZ university of applied sciences. Over more than 15 years he has been working in several fields of industry like maintenance, logistics and project management and gained a broad experience. His latest projects are focused on shopfloor automation like MES implementation and bringing industry 4.0 developments in practice.

Blaz Fortuna is the founder and CEO of extrakt.AI, a startup developing artificial intelligence-based solutions for process automation, and senior researcher at Jozef Stefan Institute. He is the initiator and primary contributor to QMiner, the open-source data analytics platform for processing large-scale real-time streams containing structured and unstructured data, and co-contributor to Event Registry. He did his Ph.D. at Jožef Stefan Institute. He was a research consultant for Bloomberg L.P., a Marie Curie Fellow at Stanford University, a postdoc at IBCN (Ghent University, Belgium), and the project manager for the XLike project.

Prof. Dr. Dunja Mladenić http://ailab.ijs.si/dunja_mladenic/ works as a researcher and a project leader at Jožef Stefan Institute, Slovenia, leading Artificial Intelligence Department and teaching at Jožef Stefan International Postgraduate School, the University of Ljubljana and the University of Zagreb. She has extensive research experience in studying and developing Machine Learning, Big Data/Text Mining, the Internet of Things, Data Science, Semantic Technology techniques, and their application to real-world problems. She has published papers in refereed journals and conferences, co-edited several books, served on program committees of international conferences, and organised international events. She serves as a project evaluator of proposals for the European Commission and USA National Science Foundation. From 2013 to 2017, she served on the Institute's Scientific Council as a vice president (2015-2017). She serves on the Executive Board of Slovenian Artificial Intelligence Society SLAIS (as a president of SLAIS (2010-2014)) and on the Advisory board of ACM Slovenija.

Data availability statement

The data is not available due to restrictions.

ORCID

Patrik Zajec http://orcid.org/0000-0002-6630-3106
Jože M. Rožanec http://orcid.org/0000-0002-3665-639X
Blaž Fortuna http://orcid.org/0000-0002-8585-9388
Dunja Mladenić http://orcid.org/0000-0002-0360-6505

References

Abd Al Rahman, M., and Alireza Mousavi. 2020. "A Review and Analysis of Automatic Optical Inspection and Quality Monitoring Methods in Electronics Industry." *IEEE Access* 8:183192–183271. https://doi.org/10.1109/Access.6287639.

Aggour, Kareem S., Vipul K. Gupta, Daniel Ruscitto, Leonardo Ajdelsztajn, Xiao Bian, Kristen H. Brosnan, Natarajan Chennimalai Kumar, and Rajkumar K. 2019. "Artificial Intelligence/machine Learning in Manufacturing and Inspection: A GE Perspective." *MRS Bulletin* 44 (7): 545–558. https://doi.org/10.1557/mrs.2019.157.

Beltrán-González, Carlos, Matteo Bustreo, and Alessio Del Bue. 2020. "External and Internal Quality Inspection of Aerospace Components." In *2020 IEEE 7th International Workshop on Metrology for AeroSpace (MetroAeroSpace)*, 351–355. IEEE.

Benbarrad, Tajeddine, Marouane Salhaoui, Soukaina Bakhat Kenitar, and Mounir Arioua. 2021. "Intelligent Machine Vision Model for Defective Product Inspection Based on Machine Learning." *Journal of Sensor and Actuator Networks* 10 (1): 7. https://doi.org/10.3390/jsan10010007.

Bergmann, Paul, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. "MVTec AD–A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9592–9600.

Bradley, Andrew P. 1997. "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms." *Pattern Recognition* 30 (7): 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Cai, Qi, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. 2018. "Memory Matching Networks for One-Shot Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4080–4088.

Caron, Mathilde, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. "Emerging Properties in Self-Supervised Vision Transformers." In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 9630–9640.

Chin, Roland T., and Charles A. Harlow. 1982. "Automated Visual Inspection: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*6): 557–573. https://doi.org/10.1109/TPAMI.1982.4767309.

Chouchene, Amal, Adriana Carvalho, Tânia M. Lima, Fernando Charrua-Santos, Gerardo J. Osório, and Walid Barhoumi. 2020. "Artificial Intelligence for Product Quality Inspection Toward Smart Industries: Quality Control of Vehicle Non-Conformities." In *2020 9th International Conference on Industrial Technology and Management (ICITM)*, 127–131. IEEE.

Ciravegna, Gabriele, Frédéric Precioso, Alessandro Betti, Kevin Mottin, and Marco Gori. 2023. "Knowledge-Driven Active Learning." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 38–54. Springer.

Cullinane, Sarah-Jane, Janine Bosak, Patrick C Flood, and Evangelia Demerouti. 2013. "Job Design Under Lean Manufacturing and Its Impact on Employee Outcomes." *Organizational Psychology Review* 3 (1): 41–61. https://doi.org/10.1177/2041386612456412.

Czimmermann, Tamás, Gastone Ciuti, Mario Milazzo, Marcello Chiurazzi, Stefano Roccella, Calogero Maria Oddo, and Paolo Dario. 2020. "Visual-based Defect Detection and Classification Approaches for Industrial Applications—A Survey." *Sensors* 20 (5): 1459. https://doi.org/10.3390/s20051459.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. "An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale." In *International Conference on Learning Representations*, https://openreview.net/forum?id = YicbFdNTTy.

Duan, Guifang, Hongcui Wang, Zhenyu Liu, and Yen-Wei Chen. 2012. "A Machine Learning-based Framework for Automatic Visual Inspection of Microdrill Bits in PCB Production." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42 (6): 1679–1689. https://doi.org/10.1109/TSMCC.2012.2216260.

EC2. n.d. "European Commission, Enabling Technologies for Industry 5.0, Results of a Workshop with Europe's Technology Leaders." https://op.europa.eu/en/publication-detail/-/publication/8e5de100-2a1c-11eb-9d7e-01aa75ed71a1/language-en. September 2020.

Ghai, Bhavya, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. "Explainable Active Learning (xal) Toward Ai Explanations As Interfaces for Machine Teachers." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW3): 1–28. https://doi.org/10.1145/3432934.

Gobert, Christian, Edward W. Reutzel, Jan Petrich, Abdalla R. Nassar, and Shashi Phoha. 2018. "Application of Supervised Machine Learning for Defect Detection During Metallic Powder Bed Fusion Additive Manufacturing Using High Resolution Imaging." *Additive Manufacturing* 21:517–528. https://doi.org/10.1016/j.addma.2018.04.005.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, Shell Xu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M. Hospedales. 2022a. "Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference." In *CVPR*.

Hu, Shell Xu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M. Hospedales. 2022b. "Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9068–9077.

Huang, Yiming, Di Wu, Zhifen Zhang, Huabin Chen, and Shanben Chen. 2017. "EMD-based Pulsed TIG Welding Process Porosity Defect Detection and Defect Diagnosis Using GA-SVM." *Journal of Materials Processing Technology* 239:92–102. https://doi.org/10.1016/j.jmatprotec.2016.07.015.

Jia, Hongbin, Yi Lu Murphey, Jinajun Shi, and Tzyy-Shuh Chang. 2004. "An Intelligent Real-Time Vision System for

Surface Defect Detection." In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 3, 239–242. IEEE.

Konstantinidis, Fotios K., Nikolaos Myrillas, Konstantinos A. Tsintotas, Spyridon G. Mouroutsos, and Antonios Gasteratos. 2023. "A Technology Maturity Assessment Framework for Industry 5.0 Machine Vision Systems Based on Systematic Literature Review in Automotive Manufacturing." *International Journal of Production Research* 1–37. https://doi.org/10.1080/00207543.2023.2270588.

Križnar, Karel, Jože M. Rožanec, Blaž Fortuna, and Dunja Mladenić. 2023. "Explainable Artificial Intelligence Meets Active Learning: A Novel GradCAM-Based Active Learning Strategy." Submitted.

Kujawińska, Agnieszka, Katarzyna Vogt, and Adam Hamrol. 2016. "The Role of Human Motivation in Quality Inspection of Production Processes." In *Advances in Ergonomics of Manufacturing: Managing the Enterprise of the Future*, 569–579. Springer.

Lenka, Sambit, Vinit Parida, and Joakim Wincent. 2017. "Digitalization Capabilities As Enablers of Value Co-creation in Servitizing Firms." *Psychology and Marketing* 34 (1): 92–100. https://doi.org/10.1002/mar.2016.34.issue-1.

Lied, Lars Harald, Maria Flavia Mogos, and Daryl John Powell. 2020. "Organizational Enablers for Digitalization in Norwegian Industry." In *IFIP International Conference on Advances in Production Management Systems*, 83–90. Springer.

Lim, Kendrik Yan Hong, Pai Zheng, and Chun-Hsien Chen. 2020. "A State-of-the-art Survey of Digital Twin: Techniques, Engineering Product Lifecycle Management and Business Innovation Perspectives." *Journal of Intelligent Manufacturing* 31 (6): 1313–1337. https://doi.org/10.1007/s10845-019-01512-w.

Liqun, Wang, Wu Jiansheng, and Wu Dingjin. 2020. "Research on Vehicle Parts Defect Detection Based on Deep Learning." In *Journal of Physics: Conference Series*, Vol. 1437, 012004. IOP Publishing.

Lv, Qianwen, and Yonghong Song. 2019. "Few-shot Learning Combine Attention Mechanism-based Defect Detection in Bar Surface." *ISIJ International* 59 (6): 1089–1097. https://doi.org/10.2355/isijinternational.ISIJINT-2018-722.

Mishra, Nikhil, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. "A Simple Neural Attentive Meta-Learner." *arXiv preprint arXiv:1707.03141*.

Munkhdalai, Tsendsuren, and Hong Yu. 2017. "Meta Networks." In *International Conference on Machine Learning*, 2554–2563. PMLR.

Nahavandi, Saeid.. 2019. "Industry 5.0–A Human-centric Solution." *Sustainability* 11 (16): 4371. https://doi.org/10.3390/su11164371.

Ochal, Mateusz, Massimiliano Patacchiola, Amos Storkey, Jose Vazquez, and Sen Wang. 2021. "Few-Shot Learning with Class Imbalance." *arXiv preprint arXiv:2101.02523*.

O'Mahony, Niall, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. 2020. "Deep Learning vs. Traditional Computer Vision." In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, 128–144. Springer.

Parnami, Archit, and Minwoo Lee. 2022. "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning." *arXiv preprint arXiv:2203.04291*.

Pouyanfar, Samira, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S. Iyengar. 2018. "A Survey on Deep Learning: Algorithms, Techniques, and Applications." *ACM Computing Surveys (CSUR)* 51 (5): 1–36. https://doi.org/10.1145/3234150.

Ren, Zhonghe, Fengzhou Fang, Ning Yan, and You Wu. 2022. "State of the Art in Defect Detection Based on Machine Vision." *International Journal of Precision Engineering and Manufacturing-Green Technology* 9 (2): 661–691. https://doi.org/10.1007/s40684-021-00343-6.

Rožanec, Jože M., Patrik Zajec, Spyros Theodoropoulos, Erik Koehorst, Blaž Fortuna, and Dunja Mladenić. 2022. "Robust Anomaly Map Assisted Multiple Defect Detection with Supervised Classification Techniques." *arXiv preprint arXiv:2212.09352*.

Rožanec, Jože M., Patrik Zajec, Elena Trajkova, Beno Šircelj, Bor Brecelj, Inna Novalija, Paulien Dam, Blaž Fortuna, and Dunja Mladenić. 2022. "Towards a Comprehensive Visual Quality Inspection for Industry 4.0." *IFAC-PapersOnLine* 55 (10): 690–695. https://doi.org/10.1016/j.ifacol.2022.09.486.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang. 2015. "Imagenet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3): 211–252. https://doi.org/10.1007/s11263-015-0816-y.

See, Judi E. 2012. "Visual Inspection: A Review of the Literature.".

Settles, Burr. 2009. "Active Learning Literature Survey.".

Shahin, Mohammad, F. Frank Chen, Ali Hosseinzadeh, Hamed Bouzary, and Awni Shahin. 2023. "Waste Reduction Via Image Classification Algorithms: Beyond the Human Eye with An AI-based Vision." *International Journal of Production Research* 1–19. https://doi.org/10.1080/00207543.2023.2225652.

Singh, Prerna. 2023. "Systematic Review of Data-centric Approaches in Artificial Intelligence and Machine Learning." *Data Science and Management* 6 (3): 144–157. https://doi.org/10.1016/j.dsm.2023.06.001.

Snell, Jake, Kevin Swersky, and Richard S. Zemel. 2017. "Prototypical Networks for Few-Shot Learning." *ArXiv abs/1703.05175*.

Takimoto, Hironori, Junya Seki, Sulfayanti F. Situju, and Akihiro Kanagawa. 2022. "Anomaly Detection Using Siamese Network with Attention Mechanism for Few-shot Learning." *Applied Artificial Intelligence* 36 (1): 2094885. https://doi.org/10.1080/08839514.2022.2094885.

Tortorella, Guilherme Luz, Michel J. Anzanello, Flavio S. Fogliatto, Jiju Antony, and Daniel Nascimento. 2023. "Effect of Industry 4.0 Technologies Adoption on the Learning Process of Workers in a Quality Inspection Operation." *International Journal of Production Research* 61 (22): 7592–7607. https://doi.org/10.1080/00207543.2022.2153943.

Triantafillou, Eleni, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, et al. 2019. "Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples." *arXiv preprint arXiv:1903.03096*.

Villalba-Diez, Javier, Daniel Schmidt, Roman Gevers, Joaquín Ordieres-Meré, Martin Buchwitz, and Wanja Wellbrock. 2019. "Deep Learning for Industrial Computer Vision Quality Control in the Printing Industry 4.0." *Sensors* 19 (18): 3987. https://doi.org/10.3390/s19183987.

Wang, Haohan, Zhuoling Li, and Haoqian Wang. 2021. "Few-shot Steel Surface Defect Detection." *IEEE Transactions on Instrumentation and Measurement* 71:1–12.

Wang, Yisen, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. 2018. "Iterative Learning with Open-Set Noisy Labels." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8688–8696.

Wang, Yaqing, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. "Generalizing From a Few Examples: A Survey on Few-shot Learning." *ACM Computing Surveys (CSUR)* 53 (3): 1–34. https://doi.org/10.1145/3386252.

Wu, Ke, Jie Tan, Jingwei Li, and Chengbao Liu. 2021. "Few-Shot Learning Approach for 3D Defect Detection in Lithium Battery." In *Journal of Physics: Conference Series*, Vol. 1884, 012024. IOP Publishing.

Xu, Jiancheng, and Jialei Ma. 2022. "Auto Parts Defect Detection Based on Few-Shot Learning." In *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*, 943–946. IEEE.

Yang, Jing, Shaobo Li, Zheng Wang, Hao Dong, Jun Wang, and Shihao Tang. 2020. "Using Deep Learning to Detect Defects in Manufacturing: a Comprehensive Survey and Current Challenges." *Materials* 13 (24): 5755. https://doi.org/10.3390/ma13245755.

Yu, Jianbo, Yanshu Wang, Qingfeng Li, Hao Li, Mingyan Ma, and Peilun Liu. 2023. "Cascaded Adaptive Global Localisation Network for Steel Defect Detection." *International Journal of Production Research* 1–18. https://doi.org/10.1080/00207543.2023.2281664.

Zavrtanik, Vitjan, Matej Kristan, and Danijel Skočaj. 2021. "DRAEM-A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.

Zhan, Zhu, Jinfeng Zhou, and Bugao Xu. 2022. "Fabric Defect Classification Using Prototypical Network of Few-shot Learning Algorithm." *Computers in Industry* 138:103628. https://doi.org/10.1016/j.compind.2022.103628.

Zhang, Shen, Fei Ye, Bingnan Wang, and Thomas G. Habetler. 2020. "Few-Shot Bearing Anomaly Detection via Model-Agnostic Meta-Learning." In *2020 23rd International Conference on Electrical Machines and Systems (ICEMS)*, 1341–1346. IEEE.

Zhao, Zetian, Bingtao Hu, Yixiong Feng, Bin Zhao, Chen Yang, Zhaoxi Hong, and Jianrong Tan. 2023. "Multi-surface Defect Detection for Universal Joint Bearings Via Multimodal Feature and Deep Transfer Learning." *International Journal of Production Research* 61 (13): 4402–4418. https://doi.org/10.1080/00207543.2022.2138613.