

Towards Interactive Anomaly Detection using Natural Language

Callum Rothon^{*}, Simon Keizer^{*}, Rama Doddipatla, Nina Dethlefs

Abstract When training models for visual anomaly detection, typically, a dataset is collected and then annotated offline. Even if collecting raw data is relatively cheap, annotations are expensive, especially if they require human expertise. We therefore propose a novel interactive learning framework that combines active learning with natural language interaction to minimise the amount of annotated training data and allow for refined human expert feedback that may be leveraged in the learning process. In our initial experiments on wind turbine drone images, we demonstrate the effectiveness of active learning for anomaly detection when using ground truth labels, and assess the impact on learning when collecting labels from ‘experts’ versus ‘non-experts’ using our dialogue system. In addition to anomaly labels with confidence scores, we collect and analyse natural language explanations, which may be used to improve both anomaly detection performance and explainability.

1 Introduction

Deep learning (DL) models for anomaly detection in images rely on large amounts of training data, which requires labour intensive human annotation. To train expert systems for specialist industrial tasks, domain expertise is necessary, which incurs a high cost for labour. User friendly programs, such as VGG Image Annotator (VIA) [5, 6], have streamlined the task as much as possible, allowing users to click and drag bounding boxes and select classes, but the task remains time intensive, and puts a heavy burden on the annotator.

Callum Rothon^{*} and Nina Dethlefs
University of Hull, United Kingdom e-mail: {c.rothon, n.dethlefs}@hull.ac.uk

Simon Keizer^{*} and Rama Doddipatla
Toshiba Europe Limited, Cambridge Research Laboratory, United Kingdom e-mail: {simon.keizer, rama.doddipatla}@toshiba.eu

^{*} Equal contribution.

Active and Interactive Machine Learning (AL and IML) are human-in-the-loop (HITL) techniques [23], which aim to make training more efficient and adaptive by involving human experts more closely in the training process. Both have been shown to be successful in object detection [17, 13] and anomaly detection tasks [36].

Our aim is to take this one step further by introducing a dialogue system to support natural language interaction between the learning agent and the human expert. This has the potential to both improve the user experience and explainability of the system, as well as offer a pathway to leverage pre-trained language models in the anomaly detection task, taking inspiration from [24, 25]. Specifically, we propose a methodology for interactively training an anomaly detection model, using a dialogue system to collect labels from a human expert. We validate the system on a task from the wind energy industry, and simultaneously use the dialogue system to elicit natural language explanations for the provided labels.

Our main contributions are as follows:

1. We propose a novel task on interactive learning for anomaly detection that features a dialogue system for querying human oracles.
2. We establish the effectiveness of our model, comparing the use of ground truth labels with interactively obtained labels from both experts and novice users.
3. We present and release a novel dataset of human-system conversations, featuring label assignments with confidence scores and natural language explanations.¹

2 Background

In this section, we briefly outline related work in visual anomaly detection, interactive machine learning, and language-assisted object detection.

2.1 Visual Anomaly Detection

Within the field of computer vision, anomaly detection has been an area of rapid development, with applications in medicine [39], manufacturing quality control [14, 42, 12, 3], and other areas [30, 43, 2]. Deep CNN (DCNN) based methods [41, 7, 14, 33, 16, 18, 35], have largely superseded other methods for anomaly detection, due to their superior performance and ability to be trained on relatively little in-domain data by fine-tuning pre-trained models. However, training DCNN-based anomaly detectors still requires large amounts of annotated data, which is labour-intensive and expensive. To further reduce the need for annotated data, AL and IML

¹ Raw and cleaned human-system conversations can be found in our Kaggle repository at <https://www.kaggle.com/datasets/54ceee996fc19201769127077d0f307f1cf3ec93d0cfdc7bf120e253963d4f96>.

approaches for anomaly detection have been proposed [40, 2, 43], and studies have trialled the use of NLP tools to generate ground truth from existing descriptions [39, 25]. To our knowledge, these approaches have not yet been combined.

2.2 Active/Interactive Machine Learning

AL and IML are Human-in-the-Loop methods that aim to involve the human expert more closely in the learning process, thereby reducing the amount of annotation data, and overall make the learning process more efficient and flexible in general [23]. [4] present a structural behavioural model of a generalised IML system, and provides recommendations for setting up a successful IML framework. [28] is focused on Interactive Machine Teaching (IMT), which relies on an iterative process in which a human assists an ML system in learning a task.

AL and IML methods have also been applied with some success to object and anomaly detection tasks [38, 1, 22, 20, 21, 19, 8]. [36] compares a range of AL and ML approaches on an anomaly detection task, finding that their AL model outperforms traditional methods. [40] uses instance-level uncertainty to rank images for sampling the most valuable images for annotation. In industrial tasks, [43] presents an AL framework for the detection of weld defects on a pipeline, using value based sampling, and [2] uses an IML-trained U-net to label patches in asphalt.

2.3 Natural Language and Vision

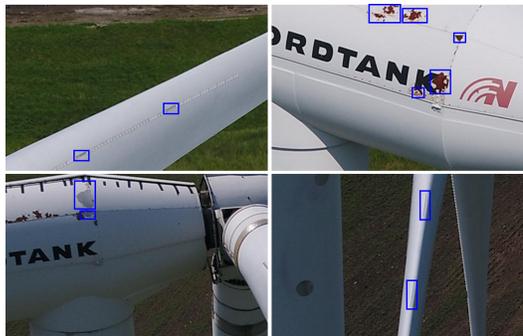
Many language-assisted object detectors extract spatial information and features from descriptions and reports. [34, 10, 15, 24, 25, 39] combine features learned from images and from text-based descriptions to improve training of object and anomaly detectors, thus requiring fewer annotated images. In contrast to the above approaches, we aim to elicit and exploit natural language descriptions from human experts in an interactive setting, envisioning a system that can generate explanations for its predictions, a user who can provide free-form feedback, and information can be clarified across multiple dialogue turns.

3 Domain and dataset

We use data from the Nordtank DTU dataset [32] to validate our model. This dataset contains UAV-gathered images of wind turbines exhibiting damage, taken over two years. 1000 images from a cropped version of the 2018 dataset were reviewed and annotated, resulting in 229 images containing 638 instances of anomaly. The anomalies fall into five main classes: corrosion (204 instances), missing paint (110

instances), missing vortex generators (26 instances), cracks (13 instances), and unknown (signs of repair, marks; 284 instances). Some examples are shown in Fig. 1.

Fig. 1 Example annotations from the cropped Nordtank DTU dataset, showing missing vortex generators (top left), corrosion (top right), missing paint (bottom left), and cracks (bottom right).



During testing, users were shown an image with a bounding box, and asked whether any anomaly was present in that bounding box, regardless of class. This simplification to binary labels was to make the task easier for the users, and will be expanded to the more complex multi-class setting in future experiments.

4 Methodology

The proposed system is based on an AL querying approach, with a dialogue system to elicit labels for the selected images from a human expert using natural language. The collected labels are then used to iteratively train a CNN-based anomaly detector. The overall IML framework is summarised in Section 4.1, whereas the dialogue system is described in more detail in Section 4.2.

4.1 Interactive Learning Framework

The IML framework is illustrated in Fig. 2. The IML Agent selects images for human annotation based on a querying strategy, and iteratively trains the anomaly detection model using the collected annotations. The proposed querying strategy consists of selecting images with low-confidence predictions by the current trained model: first, images are ranked based on the highest confidence prediction per image, and then the five lowest ranking images are selected to be reviewed by the human expert, querying up to five predicted anomalies per image.

The anomaly detection model itself is a Faster R-CNN [29] with a ResNet-50v2 [9] backbone network. This model has been shown to be effective in previous tasks

[31], and can be easily set up and fine-tuned for new applications. We verified its suitability for the current task empirically.

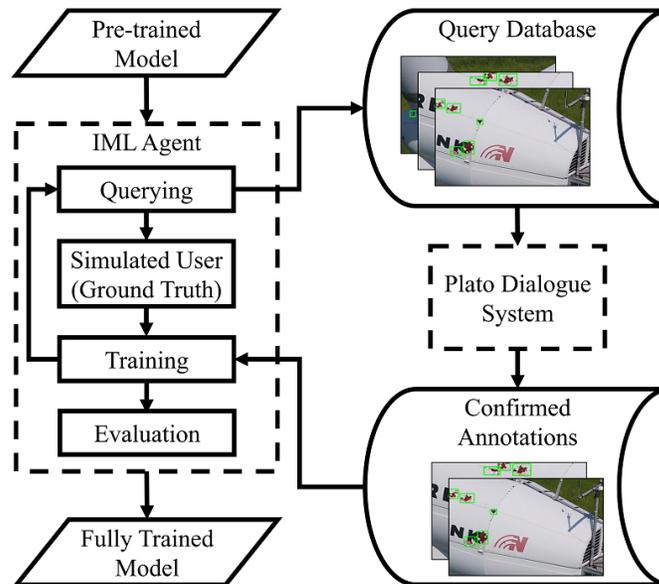


Fig. 2: A flowchart of the key components of the Interactive Machine Learning system, including the learning agent and the dialogue system.

During each stage of training, a batch of five images are selected, based on the model’s current predictions and the query strategy. These images and current predictions are then reviewed either by a simulated user (using the ground-truth label) or human user, resulting in a new batch of training data. The model is then retrained on the collected training data for 500 epochs, using a batch size of 4 images, and optimised using AdamW with a learning rate of 0.00001. Precision, Recall, and F1 scores are calculated on the test data, allowing a balanced view of the performance of the model. Predictions are considered true positives if the Intersection over Union (IoU) with the ground truth bounding box is over 0.5.

4.2 Dialogue System

The dialogue system is designed to deliver prompts and receive information in a natural conversational manner, with the goal of efficiently gathering high quality annotations from a human expert. The dialogue system was built using Uber’s Plato Dialogue system platform [27]. A flowchart of the components is shown in Fig. 3.

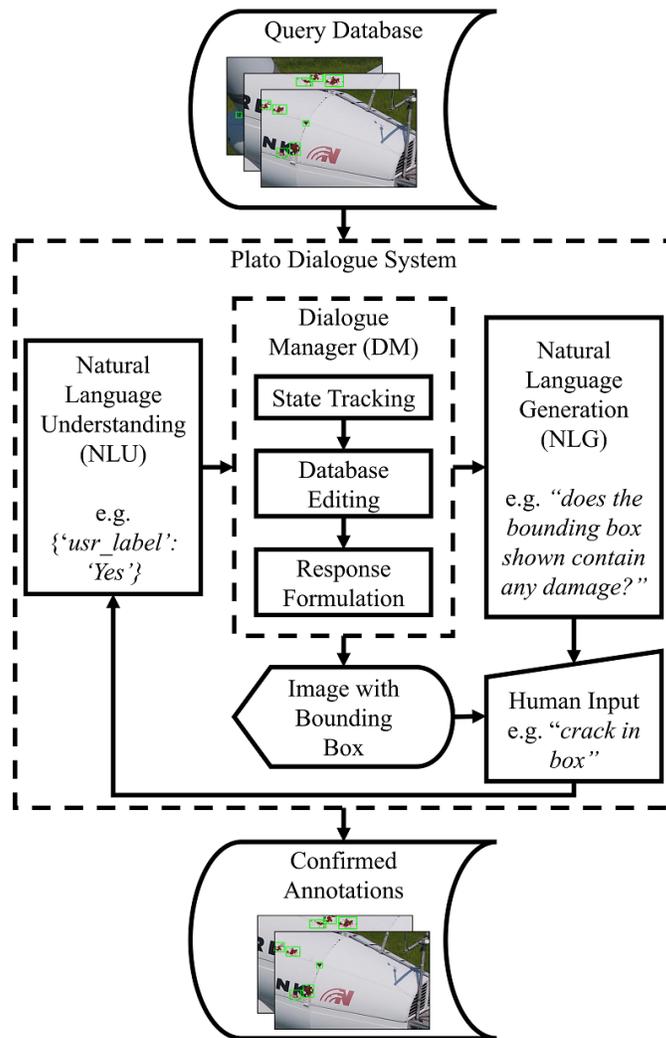


Fig. 3: Flowchart of the key components of the Plato dialogue system, including the NLU, DM, NLG, Database, and the human user.

To suit our initial experiments in this new application domain for dialogue, we developed a modular system with rule-based components for natural language understanding (NLU), dialogue management (DM), and natural language generation (NLG). The conversational data we have collected in the experiments can be used to improve these components, possibly replacing them with statistical models. As the interactive anomaly detection task grows in complexity, it will become important to use more scalable data-driven methods.

For the NLU component, we use simple pattern matching to recognise the yes/no answers used for recording binary labels for the presence/absence of damage. As we move to the multi-class scenario, this will be extended to recognising damage classes. The DM tracks the state of the conversation, for example to enable confirmation of user provided labels, and queries and updates the database that stores the images, predictions, and user provided labels. The NLG component currently uses simple rules to generate natural language realisations of the response actions selected by the DM, for example, to ask the user to label a bounding box in an image, to confirm a user provided label, or to request an explanation. At this stage, an easily interpretable rule-based dialogue system is preferable over a data-driven system, allowing us to focus on the interactive learning functionality alone. The data we collect using this system will be used in more advanced future version.

5 Validation

To validate the proposed system, we ran an AL task where users were queried on predictions made on images from the Nordtank DTU dataset [32]. Besides the human users, we also included a simulated user, which provides ground truth labels from the annotated dataset. Our experiments were focused on determining how the querying strategy and the dialogue system would impact how effectively users could provide input to the model training. The users reviewed binary predictions made by the partially trained model, confirming or negating the presence of an anomaly in a bounding box. They also provided a confidence score and an explanation for each of their judgements, which were not used in the training process, but collected for analysis and future improvements.

We recruited 20 postgraduate students to take part in our experiment, split into 10 ‘experts’ and 10 ‘non-experts’. The ‘non-experts’ were computer science students, who were only given basic instructions on how to use the dialogue system to do the task. The ‘experts’ were wind power students, who were given an additional briefing on the specific types of anomaly that can be expected in the annotation task, along with examples. Every participant interacted with the dialogue system to review the predictions made on 5 images, resulting in a pool of 50 images per group. To ensure a direct comparison between the training scenarios at each stage of training, the human users annotated the same images that were selected during active learning with the simulated user. It should be noted, however, that in a more realistic setting, images would be selected after each batch based on predictions made by the trained model in a particular scenario, and therefore the batches may be different from those in another scenario.

Following the above setup, we report results for the following 4 scenarios:

1. Training on full **ground truth** data for images which have been selected by the querying agent.
2. Training on labels provided by the **experts**.

3. Training on labels provided by the **non-experts**.
4. Training on ground truth labels for **randomly sampled** images.

We evaluate our models in terms of F1 score, rather than accuracy, due to imbalance in the dataset between the background areas with no damage and areas with damage. During the AL task, we calculate the precision, recall, and F1 scores at each batch of training to track the progress at each stage.

6 Results

We report results in terms of anomaly detection performance during training, and provide a quantitative and qualitative analysis of the interactions that provided the training data for the anomaly detection model.

6.1 Anomaly detection results

In each of the 4 conditions listed in Section 5, the models were trained for 500 epochs between batches of 5 images. The results in terms of precision, recall, and F1 score are plotted in Fig. 4.

In Fig. 4a, the model trained by active sampling from the ground truth reaches its best precision score after 6 batches of images, whereas random sampling requires 10 batches to reach equivalent precision, i.e. a reduction of 40%. The expert and non-expert trained models show more fluctuation, which indicates more noise in the training data, as may be expected when using labels from human annotators. Both eventually reach comparable precision scores to the active sampling from the ground truth, with the expert trained model outperforming the non-expert one throughout most of the training.

As shown in Fig. 4b, the ground truth active sampling and random sampling models both achieve their maximum recall after a few batches, before dropping off as the precision increases. Where the random sampling model's recall keeps getting worse, the ground truth active sampling model's recall does not deteriorate as much, and even improves significantly towards the end of the training. In contrast, the expert and non-expert trained models show a more gradual increase in recall, both eventually reaching similar levels to the random sampling model. As shown in Fig. 4c, the active sampling from the ground truth trains rapidly, and clearly outperforms random sampling, demonstrating that the active learning strategy is effective. The expert and non-expert trained models show slower training, reaching comparable results to the random sampling by the end of the training. This shows that the effectiveness of active learning depends on the expert level of the human annotator. In a real world target scenario, the human expert can be assumed to provide higher quality labels than the recruited participants in our experiment.

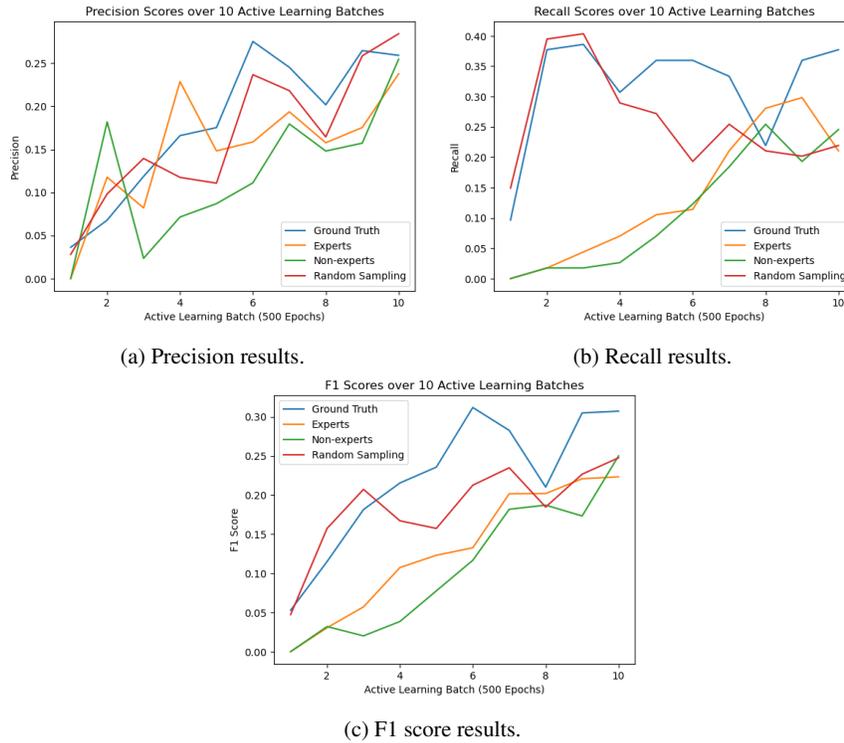


Fig. 4: Precision and recall plots of the ground truth sampling, expert, non-expert, and random sampling trained models during training.

Fluctuations seen in the training results, especially the reduced recall and F1 at batch 8 exhibited by the ground truth and random sampling are considered to be due to examples in that batch. Upon inspection of batch 8, it was found that it contained 11 instances of damage, far more than most other batches. It is thought that the significant increase in training data caused the model to not be satisfactorily trained on the new data after 500 epochs, which would explain the drop in performance.

The scores exhibited in this task are lower than in many object and defect detection tasks. It is thought that this is due to the small scale of anomalies in the images of this particular dataset, especially the crack and missing vortex generator classes. Previous work on this dataset exhibited scores in line with those achieved by our system [31]. In this study, we have focused on demonstrating the effectiveness of active learning when faced with limited data, and have collected natural language explanations that can be used to improve detection performance in a future multi-modal version of the model.

6.2 Dialogue Analysis

An example conversation from the experiment is shown in Fig. 5, contrasting expert and non-expert judgements for a selected image. The example is a false prediction, where the presence of damage was to be negated; in this case, both expert and non-expert thought there was damage, thus introducing noise in the training data that will have affected performance of the trained models.



	Expert	Non-Expert
SYSTEM	Does the bounding box shown contain any damage?	
USER	Yes	Yes
SYSTEM	Okay, confirming bounding box as showing damage, please provide an explanation.	
USER	Crack in box	Damage visible
SYSTEM	Please score how confident you are from 1-5, where 1 = Least certain, 5 = Most certain.	
USER	4	3
SYSTEM	Thank you for your feedback. Please type 'next' to see next image.	
USER	Next	Next

Fig. 5: Example interactions with an expert and a non-expert user, discussing the same image.

Expert and non-expert users took comparable amounts of time to complete the task, ranging between 2 and 4 minutes to review batches containing 15-25 predictions, corresponding around 10 seconds per predicted instance of anomaly, including explanations and confidence scores.

6.2.1 User Scores

Table 1 shows the distribution of the confidence scores provided by the expert and non-expert users. As expected, the experts were generally more confident about their judgements than the non-experts, whose confidence scores were more varied.

Table 1: Comparing confidence score frequencies between experts and non-experts.

Score	1	2	3	4	5	Mean
Experts	1	1	11	21	102	4.63
Non-experts	0	4	21	35	76	4.34

6.2.2 User Explanations

The user explanations were analysed in terms of length (number of words) and the frequency of words which occurred more than 5 times, filtering out stop words such as "the". The frequency of the lengths of explanations are shown in Fig. 6, and the frequency of recurring words are shown in Fig. 7.

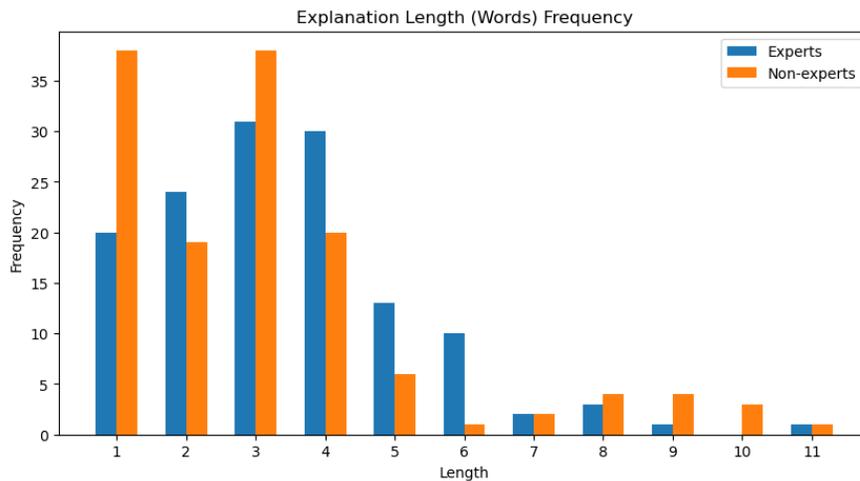


Fig. 6: A bar chart comparing occurrences of the length of user explanations, between the experts (blue) and non-experts (orange).

The results show that the user explanations were fairly short, with non-expert users providing shorter explanations than expert users. The average length of explanations was 3.53 words for experts, and 3.15 words for non-experts. This is expected, as the briefing before the experiment provided the expert users with the vocabulary to give more elaborate explanations for their judgements.

When analysing the frequency of recurring words, it was found that the experts used more specific language related to the task, such as "damage", "blade", and "box", while non-experts used less technical terms such as "part" and "shadow". This was expected, given the experts greater technical knowledge of the task. The words "box" and "damage" were found to be highly common, with utterances stating "no damage in box" being common responses to false positives.

When reviewing the user inputs qualitatively, we find that explanations associated with false positives often provide little detail, such as "no damage in box" or "cant see any damage". In some instances, the user describes a feature in the image that the model may have incorrectly classified as damage, for example "shadow in bounding box" or "part of logo or company name". Explanations associated with true positives often include a general confirmation (e.g., "damage highlighted"), which adds

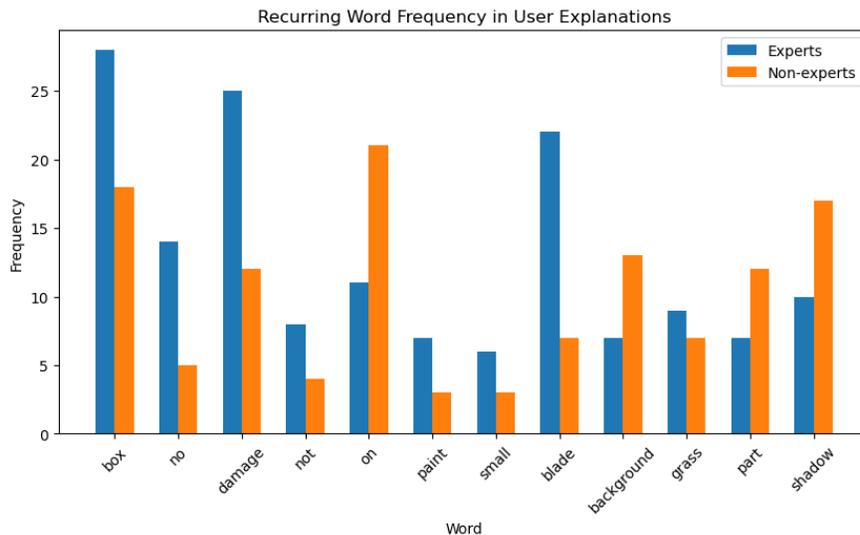


Fig. 7: A bar chart comparing occurrences of filtered words recurring more than 5 times in the user explanations, between the experts (blue) and non-experts (orange).

little value. More descriptive explanations often include the type of damage (e.g. “a couple of cracks in box”, “it looks like rust”), or may also be a more general description of what the bounding box in the image looks like (e.g. “darker coloured area”), which contains more valuable information. Explanations when the user is unsure often describe the potential damage (“could be part of the blade or crack”) or explain the uncertainty (“some imperfection but small so could be due to image quality”). If these explanations are to be used to aid in the anomaly detection task, then the more detailed and valuable explanations will provide more information. Descriptive explanations that refer to specific features could potentially be used to refine both region proposal and classification in a multi class task.

7 Discussion

In this section, we provide further discussion of the interactive learning results and the collected user explanations.

7.1 Interactive Learning

In our experiments, the model trained by active sampling from the ground truth showed the best performance overall. The results clearly demonstrate the benefit of sampling ground truth labels based on low-confidence model predictions rather than random sampling. However, this setting corresponds to a noise-free interactive learning setting where a human expert provides perfect annotations.

In our setting with recruited graduate students, grouped into ‘experts’ and ‘non-experts’, some noise in the user-provided labels was anticipated. Indeed, we found that the experts showed only 87.5% agreement with the ground truth, and the non-experts even lower, at 77.8%. This level of noise clearly affected the training results. However, we envision that end-users of the system will be fully trained wind turbine engineers that can provide much more reliable labels, which may also be further improved by using a committee of experts.

Overall, the results of the AL task on the ground-truth and user-input data showed that the proposed system has the potential to reach better performance than random sampling with fewer training examples, and with less human effort required than when annotating images by plotting bounding boxes. Training could be further improved by incorporating user provided confidence scores and explanations.

We believe that the high user confidence scores are likely due to the prevalence of false positive predictions in the earlier stages of training, which that can be easily labelled by the users as such (although Fig. 5 shows a more challenging case). The slightly higher scores from the experts indicate that their domain knowledge made them more confident compared to the non-experts.

When considering the effect of our system on annotation times, our system allows faster annotations than GUI-based click and drag annotators, with an average time of around 10 seconds per instance. Although reducing annotation times was not the focus of our experiments, [37] states annotation times of up to 35 seconds per instance when labelling through conventional means, suggesting that our approach achieves a significant reduction.

7.2 User Explanations

Explanations provided by experts were generally more informative than those provided by non-experts. The briefing given to the experts clearly led them to use more domain-specific terminology and more detail, adding value to their initial binary anomaly judgement.

The overall tendency towards shorter explanations could be partially attributed to the prevalence of false positives, especially in the earlier batches. In many such cases, the users tended to give short explanations like “nothing in box”. We also found that some users gave mostly long responses, while others were more monosyllabic, which suggests that explanation length is partially due to personal preference.

For a future version of the system that can exploit explanations in anomaly detection, we expect that more descriptive explanations will be required. This may be done by improving the experimental design or the dialogue system itself, but we are also looking into the use of multimodal Large Language Models (LLMs) such as GPT4 [26] that have emerged recently.

As a preliminary test, we have taken a small sample of images from our dataset and passed them to GPT4 with the prompt “Does this close-up of a wind turbine show any damage?”.² For example, the top right image in Fig. 1 gives the following response:

“Yes, the close-up image of the wind turbine does show some damage. The white paint on the turbine is chipped and there is visible rust. This could potentially affect the performance and lifespan of the wind turbine. It’s recommended to have it inspected and repaired by a professional to ensure its optimal operation.”

Although some of the generated descriptions are impressive, careful prompting and post-processing will be required to ensure accuracy of the responses. For example, the missing vortex generators in the top left image in Fig. 1 were not noticed:

“From the close-up image of the wind turbine blade, it appears to be in good condition with no visible damage. The blade is white in color and has a smooth surface. [...]”

Hallucinations in generated text have been identified as a significant issue in LLMs by a range of sources[11, 26], and can present plausible outputs that may be difficult to differentiate from factual responses. Therefore, having a human expert in the loop will always be essential in this kind of application.

8 Conclusion

In this paper, we have presented a novel task that uses a dialogue system to assist in data annotation for highly expert-driven tasks. We demonstrated feasibility in the domain of damage detection on wind turbine blades. Our validation study showed that the querying strategy for active learning on ground truth data was effective, and that user provided labels for active learning resulted in a comparable performance to passive learning on ground truth labels. We also collected and analysed natural language explanations from the users for their annotations, with the aim to develop a future version of the system that leverages such explanations to enhance anomaly detection performance and explainability.

Future work can extend the dialogue system to allow users to review and correct class predictions, and add previously unseen classes to the detection model, taking advantage of human ability to recognise new classes quickly. The dialogue system could be improved to better reflect how humans interact with it, using the scores and explanations gathered in the conversations to create a data-driven version, which may be more flexible and adaptable to other tasks. We will also look further into

² Using Bing enterprise, 15/16 Nov 2023.

using LLMs for data augmentation, with particular focus on image descriptions and explanations of detected anomalies.

References

1. Acuna, D., Ling, H., Kar, A., Fidler, S.: Efficient interactive annotation of segmentation datasets with polygon-rnn++. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 859–868 (2018). DOI 10.1109/CVPR.2018.00096
2. Dan, H.C., Zeng, H.F., Zhu, Z.H., Bai, G.W., Cao, W.: Methodology for interactive labeling of patched asphalt pavement images based on u-net convolutional neural network. *Sustainability* **14**(2) (2022). DOI 10.3390/su14020861. URL <https://www.mdpi.com/2071-1050/14/2/861>
3. Du, W., Shen, H., Zhang, G., Yao, X., Fu, J.: Interactive defect segmentation in x-ray images based on deep learning. *Expert Systems with Applications* **198**(116692) (2022). DOI <https://doi.org/10.1016/j.eswa.2022.116692>. URL <https://www.sciencedirect.com/science/article/pii/S0957417422001737>
4. Dudley, J.J., Kristensson, P.O.: A review of user interface design for interactive machine learning. *ACM Trans. Interact. Intell. Syst.* **8**(2) (2018). DOI 10.1145/3185517. URL <https://doi.org/10.1145/3185517>
5. Dutta, A., Gupta, A., Zisserman, A.: VGG image annotator (VIA). <http://www.robots.ox.ac.uk/vgg/software/via/> (2016)
6. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia, MM '19. ACM, New York, NY, USA (2019). DOI 10.1145/3343031.3350535. URL <https://doi.org/10.1145/3343031.3350535>
7. Feng, C., Liu, M.Y., Kao, C.C., Lee, T.Y.: Deep Active Learning for Civil Infrastructure Defect Detection and Classification, pp. 298–306 (2017). DOI 10.1061/9780784480823.036. URL <https://ascelibrary.org/doi/abs/10.1061/9780784480823.036>
8. Forte, M., Price, B.L., Xu, S.N., Pitié, F.: Getting to 99% accuracy in interactive segmentation. *CoRR* **abs/2003.07932** (2020). URL <https://arxiv.org/abs/2003.07932>
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
10. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
11. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions (2023)
12. Huang, Y.C., Hung, K.C., Lin, J.C.: Automated machine learning system for defect detection on cylindrical metal surfaces. *Sensors* **22**(24) (2022). DOI 10.3390/s22249783. URL <https://www.mdpi.com/1424-8220/22/24/9783>
13. Kellenberger, B., Marcos, D., Lobry, S., Tuia, D.: Half a percent of labels is enough: Efficient animal detection in uav imagery using deep cnns and active learning. *IEEE Transactions on Geoscience and Remote Sensing* **57**(12), 9524–9533 (2019). DOI 10.1109/TGRS.2019.2927393
14. Li, J., Su, Z., Geng, J., Yin, Y.: Real-time detection of steel strip surface defects based on improved yolo detection network. *IFAC-PapersOnLine* **51**(21), 76–81 (2018). DOI <https://doi.org/10.1016/j.ifacol.2018.09.412>. URL <https://www.sciencedirect.com/science/article/pii/S2405896318321001>. 5th IFAC Workshop on Mining, Mineral and Metal Processing MMM 2018

15. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
16. Li, Y., Han, Z., Xu, H., Liu, L., Li, X., Zhang, K.: Yolov3-lite: A lightweight crack detection network for aircraft structure based on depthwise separable convolutions. *Applied Sciences* **9**(18) (2019). DOI 10.3390/app9183781. URL <https://www.mdpi.com/2076-3417/9/18/3781>
17. Li, Z., Chen, Q., Koltun, V.: Interactive image segmentation with latent diversity. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 577–585 (2018). DOI 10.1109/CVPR.2018.00067
18. Lv, X., Duan, F., Jiang, J.J., Fu, X., Gan, L.: Deep active learning for surface defect detection. *Sensors* **20**(6) (2020). DOI 10.3390/s20061650. URL <https://www.mdpi.com/1424-8220/20/6/1650>
19. Mahadevan, S., Voigtlaender, P., Leibe, B.: Iteratively trained interactive segmentation (2018)
20. Majumder, S., Yao, A.: Scale-aware multi-level guidance for interactive instance segmentation. *CoRR abs/1812.02967* (2018). URL <http://arxiv.org/abs/1812.02967>
21. Majumder, S., Yao, A.: Content-aware multi-level guidance for interactive instance segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11,594–11,603 (2019). DOI 10.1109/CVPR.2019.01187
22. Maninis, K.K., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 616–625 (2018). DOI 10.1109/CVPR.2018.00071
23. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, : Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* **56** (2022). DOI 10.1007/s10462-022-10246-w
24. Mu, J., Liang, P., Goodman, N.: Shaping visual representations with language for few-shot classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4823–4830. Association for Computational Linguistics, Online (2020). DOI 10.18653/v1/2020.acl-main.436. URL <https://aclanthology.org/2020.acl-main.436>
25. Nishida, K., Nishida, K., Nishioka, S.: Improving few-shot image classification using machine- and user-generated natural language descriptions. In: Findings of the Association for Computational Linguistics: NAACL 2022, pp. 1421–1430. Association for Computational Linguistics, Seattle, United States (2022). DOI 10.18653/v1/2022.findings-naacl.106. URL <https://aclanthology.org/2022.findings-naacl.106>
26. OpenAI: Gpt-4 technical report (2023)
27. Papangelis, A., Namazifar, M., Khatri, C., Wang, Y.C., Molino, P., Tur, G.: Plato dialogue system: A flexible conversational ai research platform (2020)
28. Ramos, G., Meek, C., Simard, P., Suh, J., Ghorashi, S.: Interactive machine teaching: a human-centered approach to building machine-learned models. *Human-Computer Interaction* **35**(5-6), 413–451 (2020). DOI 10.1080/07370024.2020.1734931. URL <https://doi.org/10.1080/07370024.2020.1734931>
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks (2015). DOI 10.48550/ARXIV.1506.01497. URL <https://arxiv.org/abs/1506.01497>
30. Ren, Z., Fang, F., Yan, N., Wu, Y.: State of the art in defect detection based on machine vision. *International Journal of Precision Engineering and Manufacturing-Green Technology* **9**, 661–691 (2021). URL <https://api.semanticscholar.org/CorpusID:236360920>
31. Rothon, C.: Applying machine learning methods to damage detection on offshore wind turbines. In: Proceedings of the Wind Energy Science Conference (WESC) 2023 (2023)
32. Shihavuddin, A., Chen, X.: Dtu-drone inspection images of wind turbine (2018). DOI 10.17632/hd96prn3nc.2
33. Shihavuddin, A., Chen, X., Fedorov, V., Nymark Christensen, A., Andre Brogaard Riis, N., Branner, K., BJORHOLM DAHL, A., Reinhold Paulsen, R.: Wind turbine surface damage detection by deep learning aided drone inspection analysis. *Energies* **12**(4), 676 (2019)

34. Socher, R., Lin, C.C., Manning, C., Ng, A.Y.: Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 129–136 (2011)
35. Tao, X., Zhang, D., Wang, Z., Liu, X., Zhang, H., Xu, D.: Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **50**(4), 1486–1498 (2020). DOI 10.1109/TSMC.2018.2871750
36. Trajkova, E., Rožanec, J.M., Dam, P., Fortuna, B., Mladenčić, D.: Active learning for automated visual inspection of manufactured products (2021)
37. Wang, X., Wei, G., Chen, S., Liu, J.: An efficient weakly semi-supervised method for object automated annotation. *Multimedia Tools and Applications* **83**, 1–24 (2023). DOI 10.1007/s11042-023-15305-0
38. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep interactive object selection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 373–381 (2016). DOI 10.1109/CVPR.2016.47
39. Yi, P.H., Kim, T.K., Lin, C.T.: Comparison of radiologist versus natural language processing-based image annotations for deep learning system for tuberculosis screening on chest radiographs. *Clinical Imaging* **87**, 34–37 (2022). DOI <https://doi.org/10.1016/j.clinimag.2022.04.009>. URL <https://www.sciencedirect.com/science/article/pii/S0899707122001176>
40. Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q.: Multiple instance active learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5330–5339 (2021)
41. Zhang, L., Yang, F., Daniel Zhang, Y., Zhu, Y.J.: Road crack detection using deep convolutional neural network. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3708–3712 (2016). DOI 10.1109/ICIP.2016.7533052
42. Zhou, X., Fang, H., Fei, X., Shi, R., Zhang, J.: Edge-aware multi-level interactive network for salient object detection of strip steel surface defects. *IEEE Access* **9**, 149,465–149,476 (2021). DOI 10.1109/ACCESS.2021.3124814
43. Zuo, F., Liu, J., Wang, L., Qu, F., Fu, M.: Active learning-based complex pipeline weld defect detection with lightweight neural network. In: 2022 IEEE 11th Data Driven Control and Learning Systems Conference (DDCLS), pp. 712–717 (2022). DOI 10.1109/DDCLS55054.2022.9858378