The Uncertain Object: Application of Conformal Prediction to Aerial and Satellite Images

Vicky Copley VCOPLEY@DSTL.GOV.UK
Greg Finlay GFINLAY@DSTL.GOV.UK
Ben Hiett BPHIETT@DSTL.GOV.UK
Defence Science and Technology Laboratory, Porton Down, Salisbury SP4 0JQ, UK

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Satellites and airborne sensors are critical components of the modern surveillance and reconnaissance capability. A common use case involves the application of object detection models to such images in order to rapidly process the large volumes of data. This optimises use of expensive communications channel bandwidth, reduces the cognitive load on a human interpreter and accelerates the rate at which intelligence can be generated. However there is a clear need for statements of confidence in any predictions in order to provide context and enable trust in model outputs.

Our work examines the use of conformal prediction approaches to robustly quantify types of uncertainty in object detection models applied to aerial and satellite imagery for intelligence, surveillance and reconnaissance use cases. We investigate measures of detection and location uncertainty in a YOLO model and indicate how these may be leveraged conformal-wise to provide guarantees on the percentage of objects which aren't detected and the coverage of predicted bounding boxes. We find that conformal approaches provide a simple and effective means to expose the uncertainty in the outputs of an object detection model and highlight the utility of this knowledge in the intelligence setting.

Keywords: Conformal prediction, conformal risk control, distribution-free uncertainty quantification, object detection

1. Introduction

The activities of intelligence generation, surveillance and reconnaissance (ISR) form a centuries-old part of military operations (Prunckun, 2015). ISR is an integrated process which takes a tasking requirement, collects data and information to meet that requirement and translates this into an intelligence format which is useable by decision-makers (MOD, 2023). The proliferation of modern sensing technology means that intelligence is no longer scarce but even the best intelligence can be subject to a range of interpretations (Freedman, 2022). This observation has particular resonance when machine learning models are used to derive intelligence: such models typically do not come with any performance guarantees and yet they have the potential to be confidently wrong in their predictions, for example (Moon et al., 2020).

Satellites and airborne sensors are critical components of the surveillance and reconnaissance capabilities of NATO Allies[1]. One modern ISR use case applies object detection models to images from these sources in order to rapidly process large volumes of data. This

1. https://www.nato.int/cps/en/natohq/topics_111830.htm

reduces the cognitive load on a human interpreter and accelerates the rate at which intelligence is generated. However there is a need for statements of confidence in these predictions in order to provide context and enable trust in their outputs as well as potentially to direct further data collection. Intelligence analysis is itself an exercise in expert judgement under conditions of uncertainty (Dhami et al., 2015) and as part of this endeavour it is therefore vital to expose any uncertainty in an underlying machine-learning model as transparently as possible. This will allow all sources of uncertainty to be accounted for in the decision-making process.

Conformal prediction is a straightforward method of creating uncertainty intervals or sets for machine learning models which have rigorous theoretical guarantees (Vovk et al., 2022; Angelopoulos and Bates, 2022; Vovk et al., 1999). Given a notion of uncertainty, and provided that some calibration data are available, the method can be applied retrospectively to any pre-trained machine learning model with minimal assumptions. It therefore has vast potential application and has been used in areas as diverse as drug development, cough detection and medical imaging (Alvarsson et al., 2021; Ashby et al., 2022; Lu et al., 2022).

In this paper we use conformal prediction approaches to quantify sources of uncertainty in object detection models of aerial and satellite imagery for ISR use cases. Object detection has been applied to image data from earth observation satellites and airborne sensors for various purposes such as disaster response (Pi et al., 2020); tree counting (Moharram et al., 2023); and monitoring of maritime traffic (Petković et al., 2023) but to our knowledge not in an ISR setting. Previous applications of conformal approaches to object detection models include Andéol (Andéol et al., 2023) and de Grancey (de Grancey et al., 2022) who use the method to create conformalised bounding boxes in models identifying railway signals and pedestrians respectively. More recently (Timans et al., 2024) has developed a two-step conformal method which uses uncertainty in predicted class labels to inform uncertainty intervals for bounding boxes. We complement and extend this previous work to examine other sources of uncertainty and evaluate the benefit of conformal approaches to the overall ISR enterprise. The contributions of this work are as follows:

- it demonstrates the utility of conformal prediction approaches to quantify both the detection uncertainty and localisation uncertainty which exist in object detection models;
- examines the performance of two loss functions for the conformalised detection uncertainty problem and evaluates alternative nonconformity measures for detection and localisation; and
- applies the algorithms to object detection models trained on the satellite and aerial remotely sensed data which are typical of an ISR setting.

2. Data

2.1. Satellite data

Satellite data together with a pretrained object detection model were shared by a sister project. The data are multispectral RGB images from Airbus Vision-1[2] with an image size

2. Vision-1 data courtesy of Airbus Intelligence UK and UK MOD ARTEMIS programme

of 1024 x 1024 pixels and pixel size of 3.5m. The supplied object detection model, with six classes, aims to detect and identify types of maritime vessel ranging in size from small leisure craft occupying few pixels to aircraft carriers which may be over 300m long. An example Vision-1 image showing the relative size of different maritime vessels relative to each other is given in Figure 1.

The full dataset used in model development consists of more than 750 images. For conformal calibration and validation we restrict our attention to holdout data only which comprises 81 images and 1919 labelled maritime objects. Given that the model had already been trained we did not have flexibility in choice of the number of images for conformal calibration since no further images were available to our study. The limited number of calibration and validation images is consistent with potential constraints on data supply in an operational ISR setting and will help to establish if the conformal method is performant in this circumstance.

Figure 1: Example Vision-1 RGB satellite image showing relative size of different maritime vessels.

2.2. Aerial data

For aerial data we use the Aerial Floating Object (AFO) RGB dataset available from Kaggle[3][4] (Gąsienica-Józkowy et al., 2021). This contains images obtained from aerial drone video with annotated humans and other objects floating in the water. Image size is variable, ranging from 3840 x 2160 pixels to 1280 x 720 pixels. Pixel size is also variable depending on drone altitude and image distortions. Some of the floating objects are small and occupy

3. https://www.kaggle.com/datasets/jangsienicajzkowy/afo-aerial-dataset-of-floating-objects
4. used under CC licence https://creativecommons.org/licenses/by-nc-sa/3.0/igo/

only a few pixels which makes them difficult to detect, however as the images are derived from video there are many repeating similar images which may help boost object detection model performance. An example AFO image is given in Figure 2.

We make use of the complete AFO dataset which consists of more than 3500 images and 39991 labelled objects in six classes (Gąsienica-Józkowy et al., 2021). A majority subset of images was intially used to train an object detection model while 339 images containing 5392 objects were used for conformal calibration and validation.

Figure 2: Example image from Aerial Floating Object (AFO) dataset (Gąsienica-Józkowy et al., 2021).

3. Methods

3.1. Object detection

Object detection is a computer vision task which locates instances of objects in images or videos and delineates these with bounding boxes. Object detection models are commonly fitted using convolutional neural networks and many off-the-shelf model architectures are available including R-CNN (Girshick et al., 2014), Detectron2 (Wu et al., 2019) and YOLO (Redmon et al., 2016; Redmon and Farhadi, 2016, 2018). YOLO is one of the most popular architectures because of its high accuracy and overall processing speed and there are a range of YOLO architectures to suit different use cases. The YOLO v5 nano architecture is small compared to other models in the YOLO stable and a trained nano model is roughly 4MB in size (Wong et al., 2019). This requires less processing resource for both training and detection and is thus highly suitable for onboard use on lightweight satellites with limited compute which are used for ISR purposes, for example Tyche[5].

As previously noted the Vision-1 satellite data were supplied with a pre-trained model for maritime object detection. This was in YOLO v5 nano format. We trained our own YOLO v5 nano model for floating object detection on the AFO data using transfer learning and the baseline weights which are supplied in the YOLO v5 Python library (Jocher, 2020).

Although performance of an object detection model is commonly evaluated with a single number, the mean average precision metric, there are in fact at least six sources of error

5. https://www.sstl.co.uk/space-portfolio/missions-in-build/2024/project-tyche

and therefore uncertainty (Figure 3)[6]. The relative importance of the error types can vary between applications (Bolya et al., 2020). For example precise localisation may be necessary for an autonomous vehicle seeking to avoid pedestrians, while guarantees that objects are not missed may be prioritised in ISR applications. Our focus in this paper is on missed detections and localisation uncertainty.

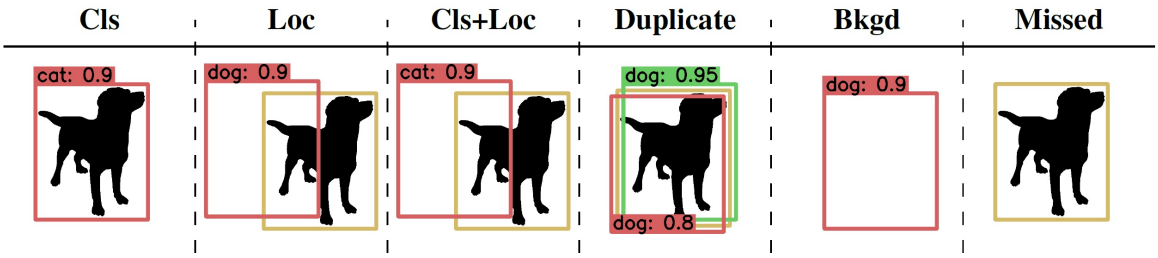


Figure 3: Possible types of error in an object detection problem. Red boxes indicate false positive detections while green boxes indicate true positives. Gold boxes represent ground truth. "Cls" = incorrect class error; "loc" = localisation error. Source: (Bolya et al., 2020).

3.2. Conformal prediction

Conformal prediction is a straightforward way to generate confidence sets or intervals for any machine learning model. Briefly, inductive conformal prediction (Papadopoulos et al., 2002) takes a pre-trained predictive model $\hat{f}$ and a holdout calibration dataset $D_{cal} = (X_i, Y_i)_{i=1}^{n}$ to compute a measure of goodness of fit between the ground truth data and model prediction $s(X, Y)$, known as a nonconformity measure (NCM) (low values are good). The choice of a nonconformity measure is a key one in the development of a conformal predictor.

Empirical quantiles of the nonconformity measure, controlled by the user-specified error level, α, may then be used to form prediction sets or ranges. Sets or ranges formed in this way are theoretically guaranteed to contain the true value with a probability of almost exactly 1-α, a property known as coverage. Formally, with X_{new} as the observed data of a new sample with Y_{new} ground truth, we obtain a set predictor $\mathcal{C}(X_{new})$ which satisfies

$$\mathbb{P}\Big(Y_{new} \in \mathcal{C}(X_{new})\Big) \geq 1 - \alpha. \tag{1}$$

Equation (1) gives the core theoretical coverage guarantee of the conformal prediction framework. The probability in Equation (1) is averaged over the randomness in the calibration and new data points and the guarantee is therefore for marginal coverage. Miscoverage is the probability that a prediction set fails to include the true value and correspondingly has probability of almost exactly α:

6. figure used under MIT licence https://github.com/dbolya/tide?tab=MIT-1-ov-file

$$\mathbb{P}\left(Y_{new} \notin \mathcal{C}(X_{new})\right) \leq \alpha. \tag{2}$$

A recent extension to the conformal framework is conformal risk control which allows for control of any single monotone loss function ℓ which shrinks as $\mathcal{C}$ grows and provides a theoretical guarantee of the following form:

$$\mathbb{E}\left[\ell(\mathcal{C}(X_{new}), Y_{new})\right] \leq \alpha. \tag{3}$$

Conformal risk control uses a prediction-sourced parameter λ to control the size of a prediction set (larger values of λ lead to larger prediction sets) (Angelopoulos et al., 2023). Risk control satisfying Equation (3) is achieved for an arbitrary risk level upper bound $\alpha \in (-\infty, B)$ by picking $\hat{\lambda}$ with the algorithm

$$\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\hat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}, \tag{4}$$

where $\hat{R}(\lambda) = (\ell(\mathcal{C}_\lambda(X_1), Y_1 + ... + \ell(\mathcal{C}_\lambda(X_n), Y_n))/n$ is the empirical risk for the calibration data. (Angelopoulos et al., 2023) show that conformal prediction reduces to conformal risk control when a miscoverage loss is used. In this case $\hat{\lambda}$ is the conformal quantile, which is selected in conformal prediction as the $\lceil (n+1)(1-\alpha) \rceil / n$ sample quantile of $\{s(X_i, Y_i)\}_{i=1}^n$.

In an object detection application theoretical guarantees may be desired either at the marginal level, considering all images and objects, or at the image-level. Conformal risk control enables the use of image-level loss functions such as false negative rate to complement the marginal coverage guarantee provided by inductive conformal prediction. Further details of the conformal framework and some of its possible applications are given in (Angelopoulos and Bates, 2022; Vovk et al., 2022).

4. Experiments

Our experiments seek to demonstrate the utility of conformal prediction approaches to quantify both the detection uncertainty and localisation uncertainty which exist in object detection models. We present two different sets of experiments:

- Detection experiments. These use the framework of conformal risk control to quantify detection uncertainty in the Vision-1 and AFO object detection models. Two loss functions are examined within this framework: marginal miscoverage and false negative rate. For each loss function two alternative nonconformity measures are also evaluated.

- Localisation experiments. These apply inductive conformal prediction to examine the effect of different nonconformity measures on the accuracy and efficiency of conformalised bounding boxes predicted by the Vision-1 object detection model.

4.1. Detections

Considering detection uncertainty, possible choices of nonconformity measure available in a YOLO v5 model include *(1 - object score)* and *(1 - confidence score)*. These are predicted directly from the underlying neural network and trained model weights and are available on a per-box basis. Object score is the model-calculated likelihood that there is an object of interest contained within a bounding box, while confidence score is equal to the object score multiplied by the score of the most likely class within that box. We test both of these nonconformity measures in our detection experiments.

In inference mode YOLO initally predicts thousands of boxes per image and uses a process of non-maximum suppression (NMS)[7] to reduce these to a user-controlled level. The ultimate number of detections shown to the user is determined by input parameters including the maximum detections per image and threshold IOU of overlapping boxes to use in the NMS process, as well as the threshold confidence score to apply to candidate boxes. We apply conformal risk control to identify the precise values which guarantee marginal miscoverage or false negative rates at given values of α, NMS IOU threshold and maximum detections per image.

To implement our approach we first randomly split the data remaining after model training into two subsets. The first of these is used for conformal calibration while the second split is used to verify the score settings identified in the calibration phase. In calibration we take predicted boxes in an image after NMS and confidence/object score thresholding have been applied and match a single predicted box to each ground truth box using the best IOU score. We count a match as a successful detection if the IOU of the predicted box with the ground truth box exceeds 0.5. Considering in turn *(1 - object score)* and *(1 - confidence score)* for matched boxes we then use conformal risk control to find threshold values which control the marginal miscoverage across all objects in the data, and the expectation of false negative rate per image. We repeat this process across 30 random splits of the data for different levels of α and monitor for miscoverage and false negative rate control as well as detections as a multiple of true detections. The latter measure is an indicator of the relative number of predicted bounding boxes compared to ground truth objects and represents the efficiency of the conformal guarantee. This determines how useful it might be to an intelligence analyst since a large surfeit of predicted bounding boxes demonstrates a level of uncertainty which may be too great for the decision problem at hand.

A summary of our experimental design is given in Table 1. Different levels of α were examined for the two models owing to their respective performance at the detection task: the Vision-1 model is much less accurate overall than the AFO model and cannot provide performance guarantees at lower levels of α without returning an unhelpfully large number of predicted boxes. For the same reason we also do not consider lower values of α such as 0.05 for either model. For all experiments we use an IOU threshold of 0.6 for the NMS process and specify maximum detections per image at an arbitrarily large 300,000 in order to avoid premature loss of predicted boxes. Exploratory analysis showed that the precise value of the NMS IOU threshold does not matter but it is important that it is held constant across calibration and validation runs to satisfy the exchangeability assumption of the conformal

7. https://pytorch.org/vision/main/generated/torchvision.ops.nms.html

method. A value of 0.6 was chosen for this threshold in order to accommodate overlapping boxes in the AFO model case.

Table 1: Summary of experiments for detections. These examine two nonconformity measures (NCM); two loss functions and an efficiency measure as performance metrics; and contrasting values of α for each model.

NCM	**Performance metrics**	**Model**	α
1 - object score	Miscoverage	Vision-1	0.5
			0.4
	False negative rate		0.3
1 - confidence score		AFO	0.3
	Detection multiple		0.2
			0.1

4.2. Localisation

A significant source of uncertainty in the object detection use case is in the location of the objects. By choosing a nonconformity measure based on the coordinates of the ground truth and predicted bounding boxes, we can use conformal prediction to provide performance guarantees on probabilities that predicted bounding boxes will fully contain the entirety of the detected object.

To do this we take the approach of (de Grancey et al., 2022) in which the nonconformity measures are set to be the distance between the ground truth coordinates for xmin, xmax, ymin and ymax and the equivalent coordinates of the predicted bounding boxes. We let $k = 1, ..., n_{box}$ index every ground truth box, irrespective of image. The ground truth coordinates of the k-th box are $Y^k = (x^k_{min}, y^k_{min}, x^k_{max}, y^k_{max})$ while $\hat{Y}^k = (\hat{x}^k_{min}, \hat{y}^k_{min}, \hat{x}^k_{max}, \hat{y}^k_{max})$ is its prediction. The nonconformity measure is defined as

$$R_k = \left(\hat{x}^k_{min} - x^k_{min}, \hat{y}^k_{min} - y^k_{min}, x^k_{max} - \hat{x}^k_{max}, y^k_{max} - \hat{y}^k_{max} \right). \tag{5}$$

Before this can be done, the correct set of ground truth bounding boxes and predictions must be paired up. In our case, as in (de Grancey et al., 2022), this was done by calculating the IOU of each prediction with each ground truth, and then applying the Hungarian matching algorithm (Kuhn, 1955) to find the minimum difference matches. Once each box has been correctly paired to its prediction, the nonconformity measures can be calculated. This leads to a set of 4 nonconformity measures and therefore a set of quantiles rather than the traditional single quantile, thus a Bonferroni correction to account for multiple comparisons must be applied here. We will refer to this set of 4 quantiles as the bounding box difference quantiles from here on. We also examine a more conservative approach to utilising these quantiles, in which rather than apply 4 different quantiles, we pick the largest difference between ground truth and prediction for each set of quantiles and apply this to

every coordinate. This method will produce larger bounding boxes on average, but will be more likely to fully cover the ground truth as a result.

Once the bounding box difference quantiles have been calculated, these are combined with future predictions to produce new conformalised bounding boxes, which have a coverage guarantee. We calculate this for a range of α values and examine the effect on coverage, IOU and the average area of bounding boxes pre- and post-conformal to evaluate the performance cost of applying conformal for localisation.

5. Results

5.1. Detections

Results of experiments for miscoverage and false negative rate are given in Table 2 and Table 3 respectively. The tables show that risk control is achieved in all cases as the mean values of miscoverage and false negative rate are less than their associated values of α.

Table 2: Results for miscoverage experiments. (NCM - nonconformity measure.)

			Miscoverage		**Detection multiple**	
NCM	**Model**	α	Mean	Std dev	Mean	Std dev
1- object score	Vision-1	0.5	0.49	0.091	1.69	0.402
		0.4	0.40	0.082	553.8	3020
		0.3	0.25	0.079	12 050	6269
	AFO	0.3	0.28	0.044	1.01	0.041
		0.2	0.19	0.044	1.20	0.094
		0.1	0.08	0.020	2025	690
1- confidence score	Vision-1	0.5	0.49	0.113	1.63	0.321
		0.4	0.39	0.087	2.46	0.773
		0.3	0.26	0.087	4248	3570
	AFO	0.3	0.28	0.036	1.00	0.034
		0.2	0.19	0.040	1.20	0.092
		0.1	0.08	0.012	2169	416

For the Vision-1 model, variation in miscoverage across experiments is slightly smaller with object score compared to confidence score (for example standard deviation of 0.091 compared to 0.113 at α=0.5 (Table 2)). However the opposite is true in the false negative rate experiments where variation in mean false negative rate across experiments is greater for object score compared to confidence score (Table 3). *(1 - confidence score)* provides a more efficient choice for λ than *(1 - object score)* for both miscoverage and false negative rate cases when measured in terms of the detection multiple (Table 2 and Table 3). Both measures become very undiscriminating of objects at lower values of α as the Vision-1 model is not performant here; although the risk control guarantees hold, detection multiples become unhelpfully large. At α=0.3 the value used to threshold the model is approaching zero so nearly all predicted boxes are returned, subject to the user-specified NMS IOU threshold

Table 3: Results for false negative rate experiments. (NCM - nonconformity measure; FNR - false negative rate.)

			FNR		**Detection multiple**	
NCM	**Model**	α	Mean	Std dev	Mean	Std dev
1 - object score	Vision-1	0.5	0.48	0.094	1.28	0.454
		0.4	0.37	0.131	1888	4892
		0.3	0.15	0.126	13 008	5367
	AFO	0.3	0.27	0.031	0.90	0.031
		0.2	0.18	0.030	1.03	0.053
		0.1	0.07	0.026	1573	1048
1 - confidence score	Vision-1	0.5	0.48	0.082	1.29	0.315
		0.4	0.37	0.097	252.3	1367
		0.3	0.14	0.111	5981	2785
	AFO	0.3	0.27	0.028	0.90	0.025
		0.2	0.19	0.028	1.01	0.055
		0.1	0.08	0.029	1274	1134

(Section 4.1). This can be seen in Figure 4 where the number of spurious detections increases as α decreases.

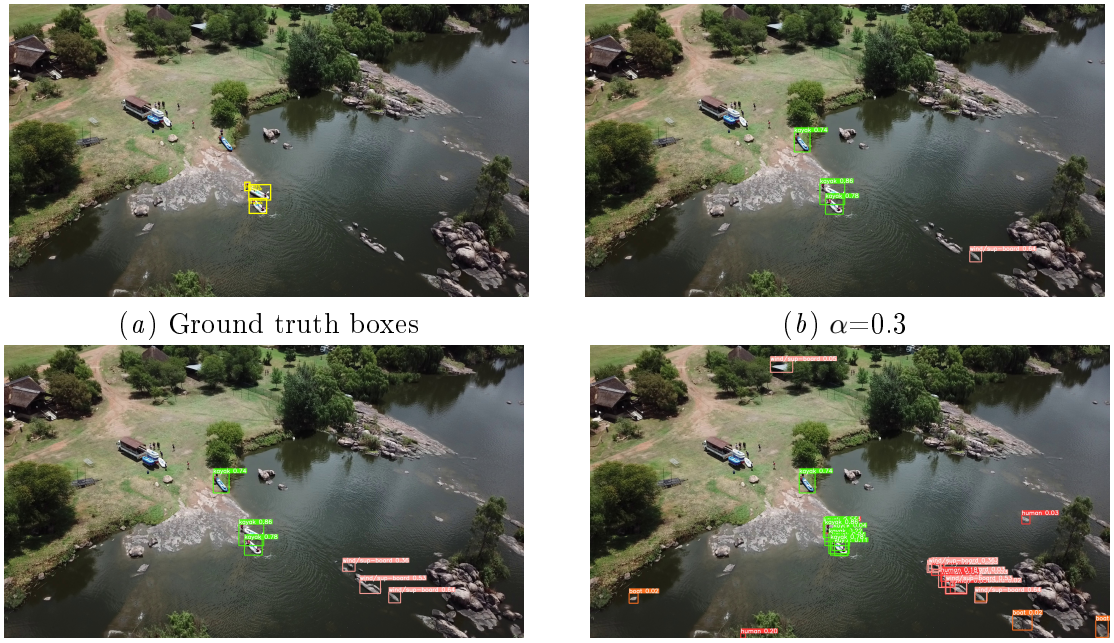
Results for the AFO model indicate no clear preference for *(1 - confidence score)* or *(1 - object score)*. Variation across miscoverage experiments is lower for *(1 - confidence score)* than it is for *(1 - object score)* (Table 2) but this does not wholly extend to false negative rate (Table 3). Risks are controlled at lower values of α than is achieved with the Vision-1 model indicating the improved predictive power of the AFO model. Excess detections proliferate at α=0.1 in a similar manner to the Vision-1 model at α=0.3 (Table 2, Table 3 and Figure 5).

5.2. Localisation

Localisation results given in Table 4 show that while the bounding box coverage guarantee is maintained at a range of values of α, the performance cost in terms of bounding box precision and size varies with α. The average IOU retained represents the average percentage change in IOU between the predicted and conformalised bounding boxes and the ground truth, a lower score indicating a larger decrease in IOU score after conformal prediction has been applied, and a score of 1 indicating no change in IOU. As α increases and the strictness of the coverage guarantee decreases, we see both the average size of the conformalised bounding boxes and the loss of IOU both decrease. When examining the effect of using the conservative bounding box quantile we see a higher level of coverage maintained overall, however with an increase in both average size of the resulting bounding box and loss of IOU score as compared to the less conservative quantile.

(*a*) Ground truth boxes (*b*) α=0.5 (*c*) α=0.3

Figure 4: Vision-1 model example false negative rate detection results using risk controlled value of confidence score at selected values of α compared with ground truth boxes in (a).



(*a*) Ground truth boxes (*b*) α=0.3

(*c*) α=0.2 (*d*) α=0.1

Figure 5: AFO model example false negative rate results using risk controlled value of confidence score at selected values of α compared with ground truth boxes in (a).

From Table 4 we can see that our coverage guarantee is in many cases is significantly higher than the required value of $(1\text{-}\alpha)$. This is most likely caused by our need for a Bonferroni correction due to having multiple quantiles for localisation, which results in generally much more conservative quantile values. Depending on the accuracy of the underlying detection algorithm, using either a more strict coverage guarantee with a lower α, or a more conservative nonconformity measure, produces larger bounding boxes in order to maintain this guarantee.

Example localisation results are shown in Figure 6 for a Vision-1 image. Figure 6 demonstrates that the performance cost of conformal prediction on localisation can be very low, as conformal boxes differ from the ground truth boxes by only a few pixels in each dimension.

Table 4: Experimental results for bounding box localisation. (NCM - nonconformity measure; IOU - intersection over union.)

NCM	α	**Coverage**	**Average IOU retained**	**Average box area change**
Bounding box	0.4	0.8222	0.6593	1.848
difference quantile	0.3	0.8444	0.6399	1.908
	0.2	0.9171	0.5688	2.217
	0.1	0.9436	0.5254	2.423
Conservative	0.4	0.9291	0.5865	2.152
bounding box	0.3	0.9291	0.5705	2.225
difference quantile	0.2	0.9675	0.4895	2.665
	0.1	0.9709	0.4567	2.849

6. Discussion

Intelligence analysis produces outputs which are ordinarily couched in terms of probabilities or uncertainty (Irwin and Mandel, 2023; Dhami et al., 2015; van der Bles et al., 2019; Friedman and Zeckhauser, 2012) and as such readily lends itself to the methods of the conformal framework which seek to expose uncertainty in black box machine learning models. We have demonstrated the successful application of conformal risk control to detections predicted by YOLO models trained using aerial and satellite data, all of which are widely used in ISR tasks. Although conformal guarantees were maintained in all cases we examined, these were less good for detections in the Vision-1 satellite maritime object model compared to the AFO model due to the respective model accuracies: higher values of the error rate α were required in the Vision-1 model case to avoid returning an impracticable number of detections. Achieving good quality object detection models with lower spatial resolution satellite data is a difficult task owing, amongst other things, to the few pixels which may be occupied by objects; relative availability of sufficient high quality training data; and greater variability of data which may span the whole planet, in contrast to aerial data which is likely to survey a few select environments at most (Van Etten, 2018; Gąsienica-Józkowy et al., 2021).

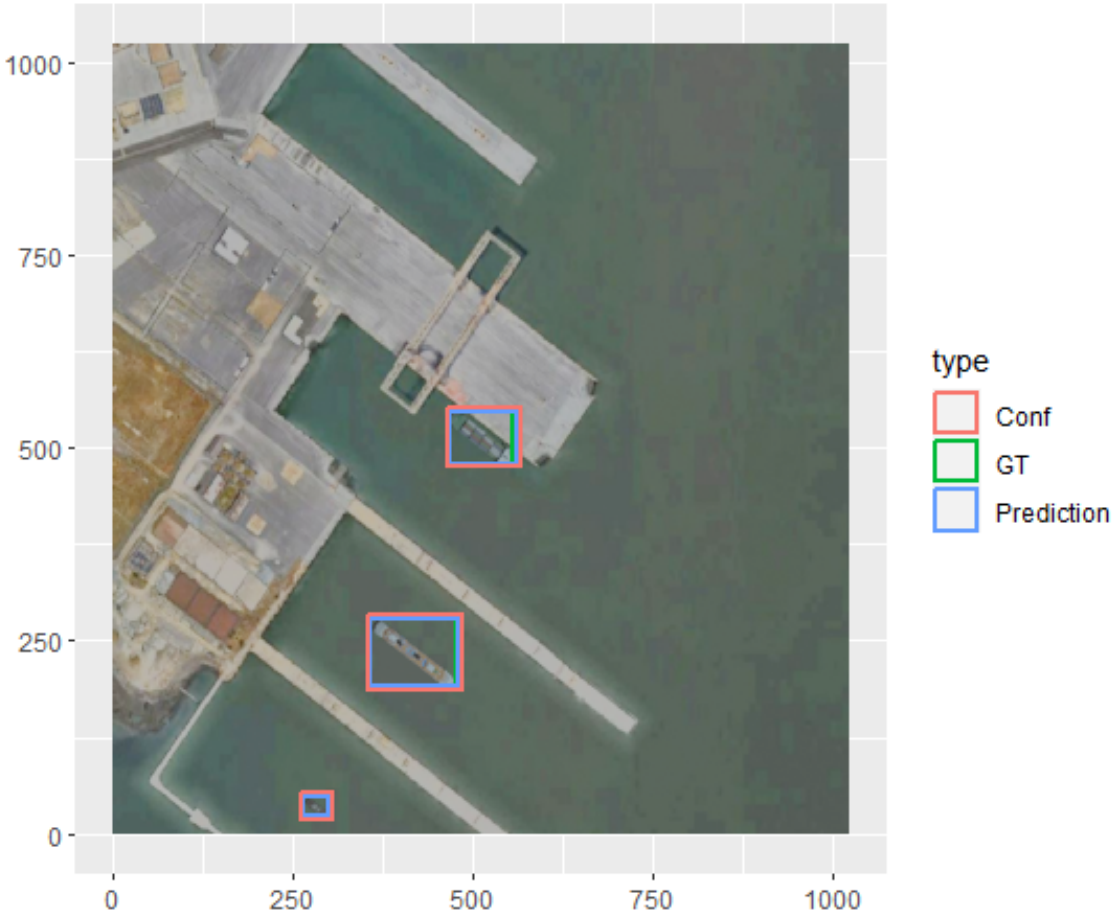


Figure 6: Vision-1 model example localisation result showing 3 objects and their associated set of bounding boxes: ground-truth (green); YOLO v5 prediction (blue); and conformalised prediction (orange).

Although it is always preferable to have a high quality trained model, the modest performance guarantees that we found for detections arising from the Vision-1 model are less germane than the ability of conformal methods to expose the underlying uncertainty in any prediction in a manner which is both flexible and versatile. The possibility to switch between loss functions in conformal risk control, for example, accommodates use cases with different objectives and offers a route by which an intelligence analyst may obtain the most appropriate measure of uncertainty for a particular intelligence assessment.

Our results also show that conformal prediction can be successfully applied to the localisation of objects in satellite images to provide coverage guarantees on the likelihood of a bounding box fully containing the target object. Although coverage was maintained in all cases, a performance cost in terms of bounding box precision and size was unavoidable. These performance costs were dependent on both values of α and choice of nonconformity measure, and as such a range of values should be examined for each use case depending on the acceptable performance loss versus coverage guarantee. In addition we found that due to the requirement for a Bonferroni correction, for localisation our coverage was often significantly more conservative than required, exceeding the value of 1-α. This suggests that in use cases in which the precision of the bounding box is valued highly, a more relaxed α may still allow a high level of coverage without commensurate loss of performance.

Conformal approaches can also be used in ISR settings to inform decisions on optimal use of the limited power, bandwidth and earth downlink opportunities which are encountered on small surveillance satellites such as Tyche. For example, images which are considered to include objects of interest, even with the 50% probability seen in our Vision-1 experiments, will be more worthy of downlink than images which do not meet this threshold. In this way the number of images for onward transmission will be substantially reduced in a theoretically rigorous manner. This has further benefit in terms of ground station processing and cognitive load on the analyst, both of which are also reduced. Ultimately, conformal methods enable greater transparency of the workings of machine learning models and this will in turn help to engender vital trust in their outputs (Bhatt et al., 2021; Radclyffe et al., 2023).

Acknowledgments

DSTL authors acknowledge the support of the UK MOD/Dstl AI Programme, Applied Verification, Validation and Vulnerabilities (AV3) Project to their research contributions. We are grateful for the comments of two anonymous reviewers which have helped to much improve the paper.

References

Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. Predicting with confidence: Using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, 2021. ISSN 0022-3549. doi: https://doi.org/10.1016/j.xphs.2020.09.055. URL https://www.sciencedirect.com/science/article/pii/S002235492030589X.

Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. 2023. doi: https://doi.org/10.48550/arXiv.2304.06052.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. 2022. doi: https://doi.org/10.48550/arXiv.2107.07511.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. 2023. doi: https://doi.org/10.48550/arXiv.2208.02814.

Alice Ashby, Julia A. Meister, Khuong An Nguyen, Zhiyuan Luo, and Werner Gentzke. Cough-based covid-19 detection with audio quality clustering and confidence measure based learning. August 2022. URL https://copa-conference.com/.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 401–413, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462571. URL https://doi.org/10.1145/3461702.3462571.

Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. 2020. doi: https://doi.org/10.48550/arXiv.2008.08115.

Florence de Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and David Vigouroux. Object Detection With Probabilistic Guarantees. In *Fifth International Workshop on Artificial Intelligence Safety Engineering (WAISE 2022)*, SAFECOMP 2022, LNCS 13415, München, Germany, September 2022. URL https://hal.science/hal-03769683.

Mandeep K Dhami, David R Mandel, Barbara A Mellers, and Philip E Tetlock. Improving intelligence analysis with decision science. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 10(6):753—757, November 2015. ISSN 1745-6916. doi: 10.1177/1745691615598511. URL https://journals.sagepub.com/doi/pdf/10.1177/1745691615598511.

Lawrence Freedman. *Command: The politics of military operations from Korea to Ukraine*. Penguin UK, 2022.

Jeffrey A. Friedman and Richard Zeckhauser. Assessing uncertainty in intelligence. *Intelligence and National Security*, 27(6):824–847, 2012. doi: 10.1080/02684527.2012.708275. URL https://doi.org/10.1080/02684527.2012.708275.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014. doi: https://doi.org/10.48550/arXiv.1311.2524.

Jan Gąsienica-Józkowy, Mateusz Knapik, and Boguslaw Cyganek. An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance. *Integrated Computer-Aided Engineering*, pages 1–15, 01 2021. doi: 10.3233/ICA-210649.

Daniel Irwin and David R. Mandel. Communicating uncertainty in national security intelligence: Expert and nonexpert interpretations of and preferences for verbal and numeric formats. *Risk Analysis*, 43(5):943–957, 2023. doi: https://doi.org/10.1111/risa.14009. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.14009.

Glenn Jocher. Yolov5 by ultralytics, 2020. URL https://github.com/ultralytics/yolov5.

Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, March 1955. doi: 10.1002/nav.3800020109.

Charles Lu, Anastasios N. Angelopoulos, and Stuart Pomerantz. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 545–554, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-16452-1_52. URL https://link.springer.com/chapter/10.1007/978-3-031-16452-1_52.

MOD. Joint doctrine note 1/23: Intelligence, surveillance and reconnaissance. Technical report, Ministry of Defence, January 2023. URL www.gov.uk/mod/dcdc.

Deema Moharram, Xuguang Yuan, and Dan Li. Tree seedlings detection and counting using a deep learning algorithm. *Applied Sciences*, 13(2), 2023. ISSN 2076-3417. doi: 10.3390/app13020895. URL https://www.mdpi.com/2076-3417/13/2/895.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7034–7044. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/moon20a.html.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36755-0. doi: 10.1007/3-540-36755-1_29. URL https://link.springer.com/chapter/10.1007/3-540-36755-1_29.

Miro Petković, Igor Vujović, Nediljko Kaštelan, and Joško Šoda. Every vessel counts: Neural network based maritime traffic counting system. *Sensors*, 23(15), 2023. ISSN 1424-8220. doi: 10.3390/s23156777. URL https://www.mdpi.com/1424-8220/23/15/6777.

Yalong Pi, Nipun D. Nath, and Amir H. Behzadan. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, 43:101009, 2020. ISSN 1474-0346. doi: https://doi.org/10.1016/j.aei.2019.101009. URL https://www.sciencedirect.com/science/article/pii/S1474034619305828.

Henry Prunckun. *How to Undertake Surveillance & Reconnaissance: From a Civilian and Military Perspective*. Pen and Sword, 2015.

Charles Radclyffe, Mafalda Ribeiro, and Robert H. Wortham. The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*, 6, 2023. ISSN 2624-8212. doi: 10.3389/frai.2023.1020592. URL https://www.frontiersin.org/articles/10.3389/frai.2023.1020592.

Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. 2016. doi: https://doi.org/10.48550/arXiv.1612.08242.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. 2018. doi: https://doi.org/10.48550/arXiv.1804.02767.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. 2016. doi: https://doi.org/10.48550/arXiv.1506.02640.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction. 2024. doi: https://doi.org/10.48550/arXiv.2403.07263.

Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana Beatriz Galvão, Lisa Zaval, and David Spiegelhalter. Communicating uncertainty about facts, numbers, and science. *Royal Society Open Science*, 6(5), May 2019. doi: https://doi.org/10.1098/rsos.181870. URL https://wrap.warwick.ac.uk/116193/.

Adam Van Etten. You only look twice: Rapid multi-scale object detection in satellite imagery. 2018. doi: https://doi.org/10.48550/arXiv.1805.09512.

V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. In *Sixteenth International Conference on Machine Learning (ICML-1999) (01/01/99)*, pages 444–453, 1999. URL https://eprints.soton.ac.uk/258960/.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Second edition, December 2022. ISBN 978-3-031-06648-1. doi: https://doi.org/10.10007/978-3-031-06649-8.

Alexander Wong, Mahmoud Famuori, Mohammad Javad Shafiee, Francis Li, Brendan Chwyl, and Jonathan Chung. Yolo nano: a highly compact you only look once convolutional neural network for object detection. 2019. doi: https://doi.org/10.48550/arXiv.1910.01271.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. 2019.