AI and Ethics (2024) 4:157–161
https://doi.org/10.1007/s43681-023-00400-7

ORIGINAL RESEARCH

Conformal prediction for trustworthy detection of railway signals

Léo Andéol[1,2] · Thomas Fel[2,3] · Florence de Grancey[4] · Luca Mossina[5]

Published online: 22 January 2024

Abstract
We present an application of conformal prediction, a form of uncertainty quantification with guarantees, to the detection of railway signals. State-of-the-art architectures are tested and the most promising one undergoes the process of conformalization, where a correction is applied to the predicted bounding boxes (i.e., to their height and width) such that they comply with a predefined probability of success. We work with a novel exploratory dataset of images taken from the perspective of a train operator, as a first step to build and validate future trustworthy machine learning models for the detection of railway signals.

Keywords Conformal prediction · Object detection · Railway signaling · Safety critical

1 Introduction

In this paper, we focus on building a trustworthy detector of railway light signals via machine learning (ML).[1] We apply uncertainty quantification (UQ) to the problem of object detection (OD) [21] via the distribution-free, non-asymptotic, and model-agnostic framework of conformal prediction (CP). CP is computationally lightweight, can be applied to any (black-box) predictor with minimal hypotheses and efforts, and has rigorous statistical guarantees.

We give a brief overview of UQ for OD and the CP method we apply to our use case. After selecting a pre-trained model among the state-of-the-art architectures, we give some insights on how UQ techniques can quantify the trustworthiness of OD models, which could be part of a critical artificial intelligence (AI) system, and their potential role in certifying such technologies for future industrial deployment.

✉ Léo Andéol
leo.andeol@math.univ-toulouse.fr

Thomas Fel
thomas_fel@brown.edu

Florence de Grancey
florence.de-grancey@irt-saintexupery.com

Luca Mossina
luca.mossina@irt-saintexupery.com

[1] Institut de Mathématiques de Toulouse, Toulouse, France

[2] SNCF, Saint-Denis, France

[3] Brown University, Providence, RI, USA

[4] Thales AVS France, Toulouse, France

[5] IRT Saint Exupéry, Toulouse, France

2 The industrial problem

AI can work as a complementary tool to enhance existing technologies, like in the automotive industry [1]. We can draw a parallel with the railway sector. While main lines (e.g., high-speed lines) already have in-cabin signaling and can be automatized [19], this is too costly to apply to the whole network. Consequently, on secondary lines, drivers can be subject to a larger cognitive load, and therefore strain, from signals and the environment. Assisting drivers with AI-based signaling recognition could facilitate the operations.

With respect to Fig. 1, our application corresponds to point (1) of their process based on multiple ML tasks: trustworthiness concerns can grow with the number of ML components. For a recent overview on the technical and regulatory challenges raised by the safety of ML systems in the railway and automotive industries, see Alecu et al. [1].

2.1 Building an exploratory dataset

Currently, there is no standard benchmark for railway signaling detection. Stemming from the dataset *FRSign* of Harb et al. [12] and their insights, our own iterations made us

[1] We refer to "safety" and "trustworthiness" intuitively; for specific examples, see [1].

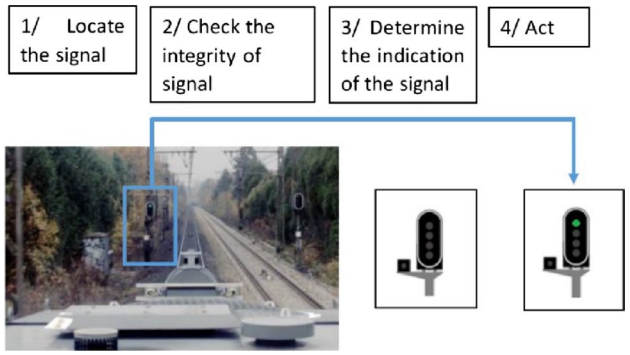

Fig. 1 Example of pipeline where an AI system acts following ML-based predictions. Source: Alecu et al. [1]

Table 1 Characteristics of our dataset

Characteristics	Quantity
Railway lines	25
Images per line (average)	55.8 ± 29.7
Images in dataset	1395
Dimensions (pixels)	1280×720
Bounding boxes (total)	2382

aware of some additional needs. Notably, our new dataset aims for images to be independent and identically distributed (i.i.d), and therefore, raw sequences cannot be considered. Moreover, our new dataset features increased variability (more railway lines, environmental and weather conditions, etc.) which could enable more accurate predictions in real-world scenarios. Finally, we generalize the task from single to multi-object detection, laying the foundations for future work in instance segmentation.

We would expect a deployable system to work whenever a human operator is successful: at night, with rain, or even when the signals are partially occluded by foliage. Within our exploratory dataset, we included different light conditions but this is far from exhaustive; these issues will be taken into account in future iterations of the dataset.

As source data, we used some video footage taken on French railway lines, with the approval of the creators.[2] To extract the samples, we acted as follows. On 25 videos of average duration about 1.5 h, we extract on average 55 images per video by running a pre-trained object detector with a low objectness threshold, and we keep a minimum interval of 5 s between detections, to prevent excessive dependence between images. We kept the images without the associated detections. We then *manually* annotated all visible railway signals. In Table 1 we report the statistics of our dataset.

[2] We would like to thank the author of the Youtube channel: https://www.youtube.com/@mika67407

3 Related works

Among the OD architecture, we point out *YOLO* [18] and its variants: they propose a one-stage detection combining convolution layers with regression and classification tasks. Howard et al. [13], with *MobileNets*, introduced the concept of depth-wise separable convolutions to reduce the number of parameters and accelerate the inference. As for the recent introduction of transformer layers, we find *DETR* [5] and *ViT* [9], among others. These networks reach state-of-the-art performances, and seem well-suited to transfer learning. Finally, *DiffusionDet* Chen et al. [6] recently formulated OD as a denoising diffusion problem, from random noisy boxes to real objects, with state-of-the-art performance.

3.1 Uncertainty quantification in object detection

UQ is an important trigger to deploy OD in transport systems. We find probabilistic OD [10], where the probability distributions of the bounding boxes and classes are predicted; Bayesian models like in Harakeh, Smart, and Waslander [11] and Bayesian approximations Li et al.deepshikha2021monte are also found in the literature. We point out the distribution-free approach of Li et al. [15]: they build probably approximately correct prediction sets via concentration inequalities, estimated via a held-out calibration set. They control the coordinates of the boxes but also the proposal and objectness scores, resulting in more and larger boxes. Finally, de Grancey et al. [7] proposed an extension of CP to OD, which will be the framework of choice in our exploratory study.

4 Conformalizing object detection

Conformal prediction (CP) [2, 20] is a family of methods to perform UQ with guarantees under the sole hypothesis of data being independent and identically distributed (or more generally exchangeable). For a specified (small) error rate $\alpha \in (0, 1)$, at inference, the CP procedure will yield prediction sets $C_\alpha(X)$ that contain the true target values Y with probability:

$$\mathbb{P}\left(Y_{new} \in C_\alpha(X_{new})\right) \geq 1 - \alpha. \tag{1}$$

This guarantee holds true, on average, over all images at inference and over many repetitions of the CP procedure. However, it is valid for any distribution of the data $\mathbb{P}_{XY}$, any sample size and any predictive model $\widehat{f}$, even if it is misspecified or a black box. The probability $1 - \alpha$ is the *nominal coverage* and the *empirical coverage* on n_{test} points is $\sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{Y_i \in C_\alpha(X_i)\}/n_{\text{test}}$.

We focus on *split* CP [14, 17], where one needs a set of *calibration* data D_{cal} drawn from the same $\mathbb{P}_{XY}$ as the test data, with no need to access the training data. At *conformalization*, we compute the *nonconformity scores*, to quantify the uncertainty on held-out data. At inference, CP adds a margin around the bounding box predicted by a pre-trained detector $\widehat{f}$.

4.1 Split conformal object detection

We follow de Grancey et al [7]. Let $k = 1, \dots, n_{box}$ index every ground-truth box in D_{cal} that was detected by $\widehat{f}$, disregarding their source image. Let $Y^k = (x^k_{\min}, y^k_{\min}, x^k_{\max}, y^k_{\max})$ be the coordinates of the k-th box and $\widehat{Y}^k = (\hat{x}^k_{\min}, \hat{y}^k_{\min}, \hat{x}^k_{\max}, \hat{y}^k_{\max})$ its prediction.

In de Grancey et al. [7], their nonconformity score, which we refer to as *additive*, is defined as:

$$R_k = (\hat{x}^k_{\min} - x^k_{\min}, \hat{y}^k_{\min} - y^k_{\min}, x^k_{\max} - \hat{x}^k_{\max}, y^k_{\max} - \hat{y}^k_{\max}). \quad (2)$$

We further define the *multiplicative* one as:

$$R_k = \Big(\frac{\hat{x}^k_{\min} - x^k_{\min}}{\widehat{w}^k}, \hat{y}^k_{\min} - y^k_{\min}\widehat{h}^k, \frac{x^k_{\max} - \hat{x}^k_{\max}}{\widehat{w}^k}, \frac{y^k_{\max} - \hat{y}^k_{\max}}{\widehat{h}^k}\Big), \quad (3)$$

where the prediction errors are scaled by the predicted width $\widehat{w}^k$ and height $\widehat{h}^k$. This is similar to Papadopoulos et al. [17], and a natural extension to de Grancey et al. [7].

Split conformal object detection goes as follows:

1. choose a nonconformity score: e.g., Equation (2) or (3);
2. Run a *pairing mechanism*, to associate predicted boxes with true boxes (see following paragraphs);
3. For every coordinate $c \in \{x_{\min}, y_{\min}, x_{\max}, y_{\max}\}$, let $\bar{R}^c = \big(R^c_k\big)_{k=1}^{n_{box}}$;
4. compute $q^c_{1-\frac{\alpha}{4}} = \lceil (n_{box}+1)(1-\frac{\alpha}{4})\rceil$-th element of the *sorted* sequence $\bar{R}^c$, $\forall c \in \{x_{\min}, y_{\min}, x_{\max}, y_{\max}\}$.

Since we work with four coordinates, for statistical reasons, we adjust the quantile order from $(1-\alpha)$ to $(1-\frac{\alpha}{4})$ using the Bonferroni correction. We conformalize *box-wise*: we want to be confident that when we detect correctly an object ("true positive"), we capture the entire ground-truth box with a frequency of at least $(1-\alpha)$, on average. During calibration (point 2. above), for all the true positive predicted boxes ($\widehat{Y}^{TP}_{cal}$), we compute the nonconformity score between the true box and the prediction. The pairing mechanism is the same as the NMS used in OD, that is, for each ground-truth bounding box, the predicted bounding boxes (not already assigned to a ground truth) are tested in decreasing order of their confidence scores. The first predicted box with an IoU above a set threshold is assigned to the ground-truth box. Note that while building ($\widehat{Y}^{TP}_{cal}$), we do not consider false negatives (due to $\widehat{f}$) which cannot be taken care of by box-wise CP.

Table 2 Comparing models via AP for two IoU threshold levels

	Average Precision	
	IoU ≥ 0.3	IoU ≥ 0.8
YOLOv5s†	0.239	0.033
YOLOv5x†	0.287	0.093
DETR resnet50	0.531	0.008
DiffusionDet	**0.839**	**0.325**

†: v5s and v5x respectively correspond to a small and a large configurations of YOLOv5

At inference, the *additive* split conformal object detection prediction box is built as:

$$\widehat{C}_\alpha(X) = \{\hat{x}_{\min} - q^{x_{\min}}_{1-\frac{\alpha}{4}}, \hat{y}_{\min} - q^{y_{\min}}_{1-\frac{\alpha}{4}}, \hat{x}_{\max} + q^{x_{\max}}_{1-\frac{\alpha}{4}}, \hat{y}_{\max} + q^{y_{\max}}_{1-\frac{\alpha}{4}}\}. \quad (4)$$

The *multiplicative* conformal prediction box is:

$$\hat{C}_\alpha(X) = \Big\{\hat{x}_{\min} - \hat{w}\cdot q^{x_{\min}}_{1-\frac{\alpha}{4}}, \hat{y}_{\min} - \hat{h}\cdot q^{y_{\min}}_{1-\frac{\alpha}{4}}, \hat{x}_{\max} + \hat{w}\cdot q^{x_{\max}}_{1-\frac{\alpha}{4}}, \hat{y}_{\max} + \hat{h}\cdot q^{y_{\max}}_{1-\frac{\alpha}{4}}\Big\}. \quad (5)$$

5 Experiments

We split our dataset into three subsets: validation, calibration, and test (respectively, of size 300, 700, 395). We compare, using the validation set, the performance of YOLOv5, DETR and DiffusionDet, pre-trained on COCO [16], as candidate base predictors $\widehat{f}$, restricting the detection to the class "traffic light". Commonly, OD models are evaluated using the average precision (AP), which is the Area Under the recall–precision Curve. AP incorporates the precision–recall trade-off, and the best value is reached when precision and recall are maximized for all objectness thresholds. In our application (Table 2), the AP for YOLOv5 is low, while DETR gives better results, and DiffusionDet is significantly superior to the others. We, therefore, use the DiffusionDet model for our conformalization. However, AP metrics alone do not give a complete picture to select OD architecture.

Fig. 2 Bounding boxes as predicted by the ML predictor (blue), their conformalized version with *additive* scores (green) and the ground truth (red). Example of images used in conformalization and testing, cropped for readability

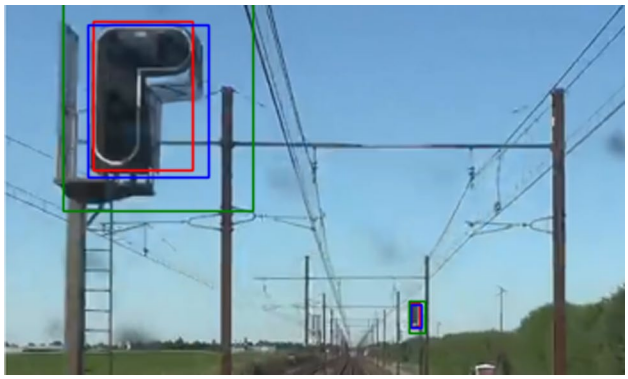

Fig. 3 Bounding boxes as predicted by the ML predictor (blue), their conformalized version with *multiplicative* scores (green) and the ground truth (red). Example of images used in conformalization and testing, cropped for readability

Table 3 Conformalization margins on DiffusionDet

	$q^{x_{\min}}_{1-\alpha}$	$q^{y_{\min}}_{1-\alpha}$	$q^{x_{\max}}_{1-\alpha}$	$q^{y_{\max}}_{1-\alpha}$	Stretch*
Additive †	4.74	7.16	6.91	5.67	$\times 2.857$
Multiplicative ‡	0.22	0.37	0.21	0.21	$\times 2.259$

*: average increase in the area of the boxes after conformalization. †: pixels; ‡: fraction of $\widehat{w}$ or $\widehat{h}$

Table 4 Empirical coverage before and after conformalization, nominal coverage: $1-\alpha = 0.90$

CP method	None	Additive	Multiplicative
Empirical coverage	0.106	0.921	0.894

This replies to: "how many true boxes, when detected, were entirely covered by the predicted box?"

5.1 Detector evaluation with box-wise conformalization

We evaluate OD performances (Table 3) by box-wise conformal prediction, reporting the estimated quantiles $q^{c}_{1-\frac{\alpha}{4}}$ at the desired risk level. A higher quantile reveals a higher uncertainty of OD predictions.In order to compare additive and multiplicative quantiles which are qualitatively different, we report the *stretch*: the ratio of the areas of the conformalized and predicted boxes. We display examples of additive and multiplicative conformalized boxes respectively on Fig. 2 and Fig. 3.

6 Results

Empirically, CP works as expected: in Table 4, we see that conformalized coverage is close to the nominal level of $(1-\alpha) = 0.9$. Remark that the guarantee holds on average over multiple repetitions, hence we cannot expect to attain the requested nominal coverage with just one dataset.

6.1 Interpretation of conformal bounding boxes

A conformalized predictor is only as good as its base predictor $\widehat{f}$. If the latter misses many ground-truth boxes, guaranteeing $(1-\alpha)$ 100% correct predictions of a few boxes will still be a small number. That is, conformalization is not a substitute for careful training and fine-tuning of a detection architecture, but a complementary tool for certification.

The interest of capturing the whole box can be operational: our ML pipeline could rely on a conservative estimation of the ground-truth to carry out a control operations (e.g., running a ML subcomponent on the detection area).

7 Conclusion and perspectives

Given the insights from this exploratory study, we plan on building and publishing an augmented version of the dataset. The objective is to have a dedicated, high-quality benchmark for the scientific community and the transport industry. As mentioned above, CP works with exchangeable data. In the long term, if trustworthy AI components are to be deployed, the UQ guarantees will need to be adapted to streams of data: this will pose a theoretical challenge and one in the construction and validation of the dataset.

So far, the underlying criterion for successful prediction has been whether the ground-truth box is *entirely* covered by the predicted box. This is strict, as having a system that guarantees to cover a big part of the truth, seems to be equally useful in practice. Bates et al. [4], with their *risk controlling prediction sets*, and Angelopoulos et al. [3], with *conformal risk control*, go in this direction. They extend the guarantee of CP to arbitrary losses, that can incorporate other operational needs.

References

1. Alecu, L., Bonnin, H., Fel, T., Gardes, L., Gerchinovitz, S., Ponsolle, L., Mamalet, F., Jenn, É., Mussot, V., Cappi, C., Delmas, K., and Lefevre, B: Can we reconcile safety objectives with machine learning performances? In ERTS (2022)
2. Angelopoulos, A. N., Bates, S: A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511 (2021)
3. Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., Schuster, T: Conformal Risk Control. arXiv:2208.02814 (2022)
4. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.: Distribution-free, risk-controlling prediction sets. J. ACM **68**(6), 1–34 (2021)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S: . End-to-end object detection with transformers. In ECCV 2020, 213–229. Springer (2020)
6. Chen, S., Sun, P., Song, Y., and Luo, P: Diffusiondet: Diffusion model for object detection. arXiv:2211.09788 (2022)
7. de Grancey, F., Adam, J.-L., Alecu, L., Gerchinovitz, S., Mamalet, F., Vigouroux, D: Object Detection with Probabilistic Guarantees: A Conformal Prediction Approach. In SAFECOMP 2022 Workshops. Springer (2022)
8. Deepshikha, K., Yelleni, S. H., Srijith, P., Mohan, C. K: Monte carlo dropblock for modelling uncertainty in object detection. arXiv:2108.03614 (2021)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
10. Hall, D., Dayoub, F., Skinner, J., Zhang, H., Miller, D., Corke, P., Carneiro, G., Angelova, A., Sünderhauf, N:Probabilistic object detection: Definition and evaluation. In Proceedings of WACV, 1031–1040 (2020)
11. Harakeh, A., Smart, M., Waslander, S. L: Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. In Proceedings of ICRA (2020)
12. Harb, J., N., Chosidow, R., Roblin, G., Potarusov, R., Hajri, H: FRSign: A Large-Scale Traffic Light Dataset for Autonomous Trains. arXiv:2002.05665 (2020)
13. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)
14. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. J. Am. Stat. Assoc. **113**(523), 1094–111 (2018)
15. Li, S., Park, S., Ji, X., Lee, I., Bastani, O:Towards PAC Multi-Object Detection and Tracking. arXiv:2204.07482 (2022)
16. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. Springer, In ECCV (2014)
17. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A: Inductive confidence machines for regression. In Proceedings of ECML, 345–356. Springer (2002)
18. Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In Proceedings of CVPR, 779–788
19. Singh, P., Dulebenets, M.A., Pasha, J., Gonzalez, E.D.R.S., Lau, Y.-Y., Kampmann, R.: Deployment of autonomous trains in rail transportation: current trends and existing challenges. IEEE Access **9**, 91427–91461 (2021)
20. Vovk, V., Gammerman, A., Shafer, G: Algorithmic Learning in a Random World. Springer, 2nd edition (2022)
21. Zhao, Z.-Q., Zheng, P., Xu, S.-T., Wu, X.: Object detection with Deep learning: a review. IEEE Trans. Neural Netw. Learn. Syst. **30**(11), 3212–32 (2019)