2022-06-16
Kasper Socha, Markus Borg
RISE Research Institutes of Sweden

Internal ML Model Test Protocol

Introduction

These results provide evidence that the ML model satisfies the ML safety requirements on the internal test data. A fundamental aspect of the argumentation is that this data set was never used in any way during the development of the ML model.

The ML model test cases provide results for both 1) the entire verification data set and 2) seven slices of the data set that are deemed particularly important. The selection of slices was motivated by either an analysis of the available technology or ethical considerations, especially from the perspective of AI fairness. Consequently, we measure the performance for the following slices of data. Identifiers in parentheses show direct connections to requirements.

S1 The entire verification data set

S2 Pedestrians close to the ego car (longitudinal distance < 50 m)(SYS-PER-REQ1, SYS-PER-REQ2)

S3 Pedestrians far from the ego car (longitudinal distance ≥ 50 m)

S4 Running pedestrians (speed ≥ 3 m/s) (SYS-ROB-REQ2)

S5 Walking pedestrians (speed > 0 m/s but < 3 m/s) (SYS-ROB-REQ2)

S6 Occluded pedestrians (entering or leaving the field of view, defined as bounding box in contact with any edge of image) (DAT-COM-REQ4)

S7 Male pedestrians (DAT-COM-REQ2)

S8 Female pedestrians (DAT-COM-REQ2)

For the ML model in SMIRK, we evaluate pedestrian detection at IoU = 0.5, which for each image means:

TP True positive: IoU ≥ 0.5

FP False positive: IoU < 0.5

FN False negative: There is a ground truth bounding box in the image, but no predicted bounding box.

TN True negative: All parts of the image with neither a ground truth nor a predicted bounding box. This output carries no meaning in our case.

Results

The total number of images in the internal test data is 139,526 (135,139 pedestrians (96.9%) and 4,387 non-pedestrians (3.1%)). Figure 1 depicts four subplots representing IoU = 0.5: A) P vs R, B) F1 vs. Conf, C) P vs. Conf, and D) R vs. Conf. Subplot A) shows that the ML model is highly accurate, i.e., the unavoidable discrimination-efficiency tradeoff of object detection is only visible in the upper right corner. Subplots B)–D) shows how P, R, and F1 vary with different Conf thresholds.

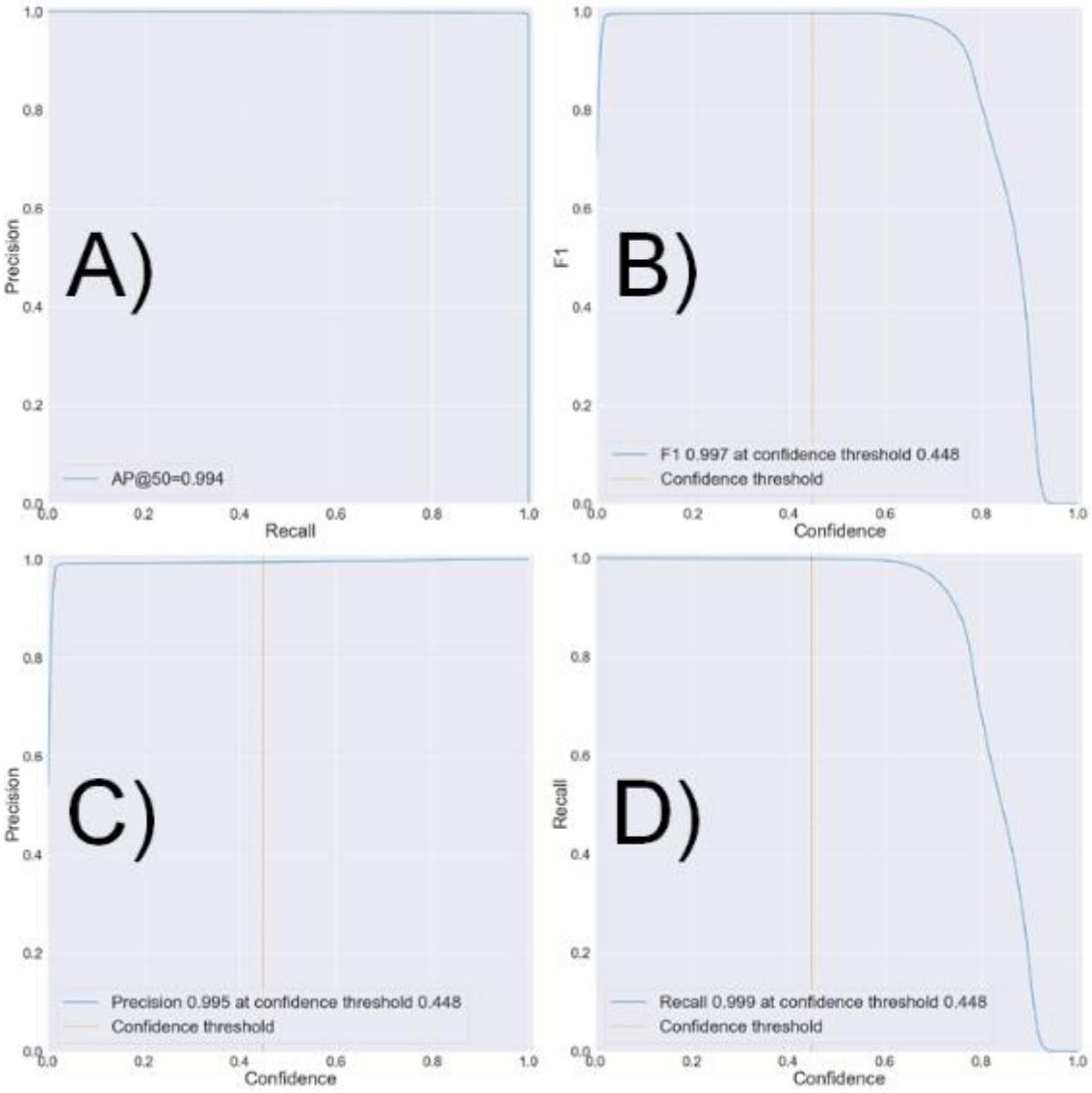

Figure 1: Evaluation of the ML model on the internal test data at IoU=0.5.

Table 1 presents further details of the accuracy of the ML model for the selected Conf threshold,organized into 1) all distances from the ego car, 2) within 80 m, and 3) within 50 m, respectively. The table also shows the effect of adding OOD detection using the autoencoder, i.e., a substantially reduced number of FPs.

Table 1: ML model accuracy on the internal test data at the Conf threshold 0.448.

Distance	Total	TP	FP	FN	P	R	F1	AP@0.5
All	139,526	134,948	711	191	0.9948	0.9986	0.9967	0.9942
+OOD		134,927	20	212	0.9999	0.9984	0.9991	0.995
$\leq$ 80 m	105,588	101,320	444	173	0.9956	0.9983	0.997	0.9948
+OOD		101,300	13	193	0.9999	0.9981	0.999	0.995
$\leq$ 50 m	61,845	57,877	186	173	0.9968	0.9970	0.9969	0.9944
+OOD		57,857	13	193	0.9998	0.9967	0.9982	0.995

Table 6 demonstrates how the ML model satisfies the performance requirements on the internal test data. First, the TP rate (95.9%) and the FN rate (0.31%) for the respective distances meet the requirements. The model's FPPI (0.42%), on the other hand, is too high to meet SYS-PER-REQ3 as we observed 444 FPs (cones outnumber spheres by 2:1). This observation reinforces the need to use a safety cage architecture, i.e., OOD detection that can reject input that does not resemble the training data. The rightmost column in Table 6 shows how the FPPI decreased to 0.012% with the autoencoder. All basic shapes were rejected, but 13 images with pedestrians led to FPs within the ODD due to too low IoU scores.

Table 2: ML model satisfaction of the performance requirements on the internal test data at the Conf threshold 0.448. R1–R4 = SYS-PER-REQ1–4.

Req.	Expected	Observed (Model)	Observed (Model+OOD)
R1	TP rate $\geq$ 93% for $\leq$ 80 m	$\frac{101,320}{105,588} = 96.0\%$	$\frac{101,300}{105,588} = 95.9\%$
R2	FN rate $\leq$ 7% for $\leq$ 50 m	$\frac{173}{61,845} = 0.28\%$	$\frac{193}{61,845} = 0.31\%$
R3	FPPI $\leq$ 0.1% for $\leq$ 80 m	$\frac{444}{105,588} = 0.42\%$	$\frac{13}{105,588} = 0.012\%$
R4	$\leq$ 3% of rolling windows contain $\geq$ 2 misses in 5 frames for $\leq$ 80 m	$\frac{216}{101,564} = 0.21\%$	$\frac{239}{101,564} = 0.24\%$

SYS-PER-REQ4 is met as the fraction of rolling windows with more than a single FN is 0.24%, i.e., ≤ 3%. Figure 2 shows the distribution of position errors in the object detection for pedestrians within 80 m of ego car, i.e., the difference between the object detection position and ESI Pro-SiVIC ground truth. The median error is 1.0 cm, the 99% percentile is 5.6 cm, and the largest observed error is 12.7 cm. Thus, we show that SYS-PER- REQ5 is satisfied for the internal test data, i.e., ≤ 50 cm position error for pedestrian detection within 80 m. Note that satisfaction of SYS-PER-REQ6, i.e., sufficient inference speed, is demonstrated as part of the system testing.

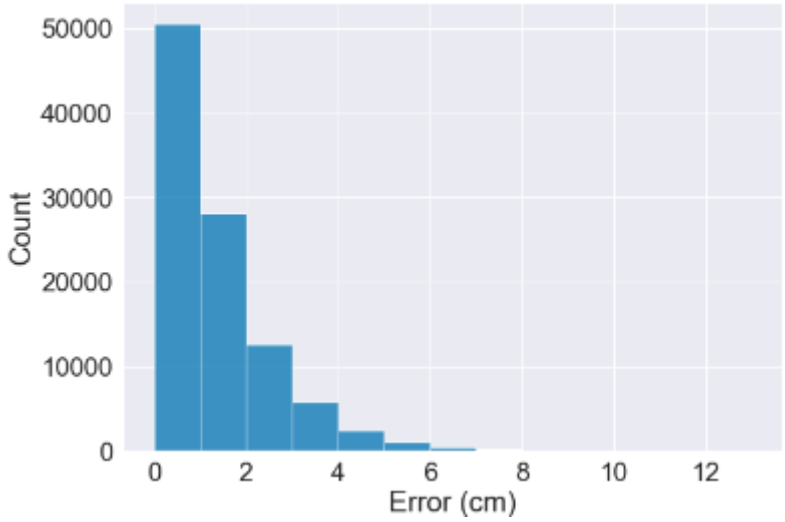

Figure 2: Distribution of position errors for the internal test data.

Table 3 presents the output of the ML model on the eight slices of internal test data S1-S8. Note that we saved the children in the ESI Pro-SiVIC object catalog for the verification data, i.e., S9 does not exist in the internal test data. Apart from the S6 slice with occlusion, the model accuracy is consistent across the slices which corroborates satisfaction of the robustness requirements on the internal test data, e.g., in relation to pose (SYS-ROB-REQ2), size (SYS-ROB-REQ2), and appearance (SYS-ROB-REQ2).

Table 3: ML model accuracy on eight slices of the internal test data. Every second rows show results for the ML model followed by OOD detection using the autoencoder.

Slice	Total	TP	FP	FN	P	R	F1	AP@0.5
S1	139,526	134,948	711	191	0.9948	0.9986	0.9967	0.9942
	+OOD	134,927	20	212	0.9999	0.9984	0.9991	0.995
S2	61,333	57,774	16	172	0.9997	0.997	0.9984	0.995
	+OOD	57,753	13	193	0.9998	0.9967	0.9982	0.995
S3	43,547	43,547	0	0	1	1	1	0.995
	+OOD	43,547	0	0	1	1	1	0.995
S4	38,786	37,804	9	48	0.9998	0.9987	0.9992	0.995
	+OOD	37,783	8	69	0.9998	0.9982	0.9990	0.995
S5	99,740	97,144	14	143	0.9999	0.9985	0.9992	0.995
	+OOD	97,144	12	143	0.9999	0.9985	0.9992	0.995
S6	778	609	16	169	0.9744	0.7823	0.8679	0.9211
	+OOD	593	13	185	0.9785	0.7618	0.8567	0.8899
S7	69,238	67,470	14	99	0.9998	0.9985	0.9992	0.995
	+OOD	67,460	11	109	0.9998	0.9984	0.9991	0.995
S8	69,288	67,479	9	91	0.9999	0.9987	0.9993	0.995
	+OOD	67,468	9	102	0.9999	0.9985	0.9992	0.995