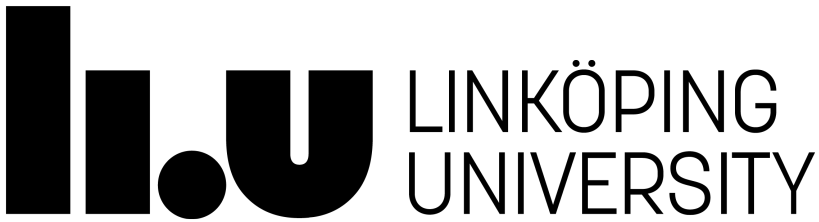
Performance of Conformal Prediction in Capturing Aleatoric Uncertainty

Misgina Tsighe Hagos, Claes Lundström

misgina@naiss.se

March 8, 2026



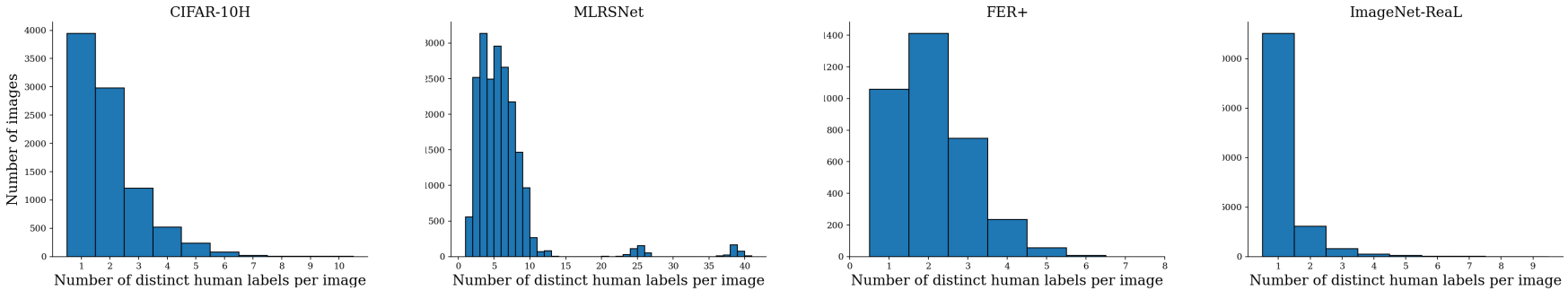
Introduction

- Conformal prediction (CP) is a framework for uncertainty quantification that generates a prediction set $\mathcal{C}(X_{\text{test}}) \subseteq \mathcal{Y}$ for a test input $(X_{\text{test}}, Y_{\text{test}})$, as opposed to conventional maximum-likelihood predictions [1, 3].
- Model uncertainty sources are categorized into aleatoric (inherent randomness) and epistemic uncertainty (lack of knowledge).
- The literature usually assume that the size of the prediction sets captures aleatoric uncertainty. However, there is no clear empirical evidence of how good conformal predictors are at it, or if they do at all.

Introduction

- We focus on aleatoric uncertainty stemming from class overlap, i.e., cases where semantically distinct classes share similar visual characteristics.
  - An instance exhibits class overlap if it collects ≥ 2 distinct labels from annotators and it is formulated as the number of distinct human labels.
  - To capture class overlap, we use datasets whose test set was annotated by multiple annotators per image.
- We assess the alignment of CP with an ambiguity that is quantified using class overlap.
  - Prediction set size vs. class overlap.
- CP approaches: the least ambiguous set-valued classifier (LAC) [2], adaptive prediction sets (APS) [3], and regularized adaptive prediction sets (RAPS) [4].

Experiments



- Four datasets that contain labels from multiple human annotators per image (ranging between five and fifty participants per image) and eight CNN models trained from scratch.
- Spearman's rank correlation coefficient to assess correlations.
- Similarity analysis between the prediction sets and human annotations using precision, recall, subset-accuracy, and Hamming loss.
- Coverage, size-stratified coverage (SSC), and mean prediction set size to evaluate the performance of the conformal predictors.

Results: correlation analysis

- Limited alignment of prediction set sizes with human perceived ambiguity as shown using correlation and similarity metrics.
- Consistently very weak correlations on the challenging tasks (97.2% of MLRSNet images have class overlap).
- Larger prediction sets lead to an improved correlation with class overlap. However, they do not ensure a better coverage.

Table: Spearman's rank correlation coefficient, $p < .001$, between prediction set sizes and class overlap.

Models	CIFAR-10H			MLRSNet			FER+			ImageNet-Real		
	LAC	APS	RAPS	LAC	APS	RAPS	LAC	APS	RAPS	LAC	APS	RAPS
ResNet18	0.131	0.231	0.231	0.045	0.064	0.064	0.365	0.325	0.326	0.320	0.316	0.316
ResNet34	0.125	0.209	0.199	0.020	0.051	0.052	0.363	0.240	0.232	0.353	0.355	0.355
ResNet50	0.144	0.215	0.219	0.038	0.049	0.050	0.364	0.302	0.280	0.361	0.364	0.364
VGG-16	0.137	0.230	0.220	-0.010	-0.016	-0.025	0.409	0.277	0.268	0.344	0.347	0.347
VGG-19	0.148	0.222	0.215	0.006	0.007	-0.003	0.390	0.265	0.252	0.352	0.349	0.349
DenseNet121	0.018	0.221	0.204	0.037	0.058	0.058	0.393	0.177	0.168	0.351	0.357	0.357
DenseNet161	0.056	0.201	0.196	0.023	0.037	0.037	0.402	0.253	0.237	0.348	0.355	0.355
MobileNet-v2	0.109	0.256	0.248	0.060	0.064	0.066	0.374	0.315	0.307	0.332	0.331	0.331
Ideal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Results: similarity analysis

- Precision and recall results are promising, though both are dependent on set size.
- Exact match or Subset accuracy is consistently low across all models and methods.

Table: Similarity between conformal predictors and human annotations on the ImageNet-ReaL dataset

Models	LAC				APS				RAPS			
	Precision	Recall	Subset accuracy	Hamming loss (↓)	Precision	Recall	Subset accuracy	Hamming loss (↓)	Precision	Recall	Subset accuracy	Hamming loss (↓)
ResNet18	0.571	0.892	0.002	0.370	0.568	0.934	0.012	0.417	0.567	0.934	0.012	0.416
ResNet34	0.670	0.885	0.001	0.481	0.631	0.930	0.008	0.480	0.629	0.931	0.008	0.478
ResNet50	0.736	0.882	0.001	0.559	0.678	0.931	0.007	0.529	0.679	0.931	0.007	0.529
VGG-16	0.628	0.887	0.002	0.434	0.604	0.933	0.010	0.453	0.603	0.934	0.010	0.452
VGG-19	0.643	0.886	0.002	0.452	0.611	0.934	0.010	0.462	0.612	0.934	0.010	0.462
DenseNet121	0.684	0.884	0.001	0.494	0.632	0.934	0.008	0.479	0.632	0.934	0.008	0.479
DenseNet161	0.763	0.876	0.001	0.593	0.705	0.921	0.005	0.552	0.702	0.923	0.005	0.548
MobileNet-v2	0.625	0.887	0.002	0.430	0.604	0.930	0.008	0.452	0.599	0.932	0.009	0.447

Results: CP performance

- All CP methods achieve better coverage of the true class.

Table: Mean ($\pm$ std) of models' coverage and SSC for each dataset.

Dataset	Accuracy	Coverage			SSC		
		LAC	APS	RAPS	LAC	APS	RAPS
CIFAR-10H	0.937 $\pm$ 0.004	0.949 $\pm$ 0.005	0.978 $\pm$ 0.003	0.977 $\pm$ 0.003	0.385 $\pm$ 0.428	0.939 $\pm$ 0.016	0.935 $\pm$ 0.031
MLRSNet	0.903 $\pm$ 0.038	0.952 $\pm$ 0.002	0.976 $\pm$ 0.012	0.979 $\pm$ 0.007	0.500 $\pm$ 0.416	0.704 $\pm$ 0.321	0.810 $\pm$ 0.158
FER+	0.832 $\pm$ 0.006	0.943 $\pm$ 0.002	0.963 $\pm$ 0.004	0.964 $\pm$ 0.004	0.891 $\pm$ 0.051	0.900 $\pm$ 0.058	0.915 $\pm$ 0.038
ImageNet-ReaL	0.759 $\pm$ 0.023	0.902 $\pm$ 0.001	0.943 $\pm$ 0.003	0.943 $\pm$ 0.003	0.234 $\pm$ 0.327	0.000 $\pm$ 0.000	0.062 $\pm$ 0.177

Results: sample outputs




Distinct human labels	airplane, bird, cat, dog	airplane, airport, buildings, cars, pavement, trees	neutral, sadness	European fire salamander, American chameleon, frilled lizard
Test label	bird	airplane	sadness	European fire salamander
Prediction sets: LAC	bird	airplane, railway station	neutral, sadness	European fire salamander, box turtle, green lizard
Prediction sets: APS	bird	airplane, railway station	neutral, sadness, disgust	European fire salamander, box turtle, green lizard, banded gecko, American chameleon
Prediction sets: RAPS	bird	airplane, railway station	neutral, sadness, disgust	European fire salamander, box turtle, green lizard, banded gecko, American chameleon

Figure: Conformal prediction set outputs for sample images from the four datasets. Columns from left to right: CIFAR-10H (leftmost), MLRSNet, FER+, and ImageNet-Real (rightmost).

Conclusion

- The correlation between conformal prediction set uncertainty and class overlap was frequently in the very weak to weak range.
- In agreement with previous literature, conformal predictors lead to a better coverage of the true class.
- Surprisingly, we find that increasing prediction set size does not translate to a better coverage.
- Our findings underscore that prediction set size should not be uncritically interpreted as a proxy for aleatoric uncertainty, particularly in the presence of inherent labeling ambiguity.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Bibliography

[1] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Vol. 29. Springer, 2005.

[2] Mauricio Sadinle, Jing Lei, and Larry Wasserman. "Least ambiguous set-valued classifiers with bounded error levels". In: *Journal of the American Statistical Association* 114.525 (2019), pp. 223–234.

[3] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. "Classification with valid and adaptive coverage". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3581–3591.

[4] Anastasios Nikolas Angelopoulos et al. "Uncertainty Sets for Image Classifiers using Conformal Prediction". In: *International Conference on Learning Representations*. 2021.

Performance of Conformal Prediction in Capturing Aleatoric Uncertainty

Misgina Tsighe Hagos, Claes Lundström

misgina@naiss.se

March 8, 2026