Beyond Memorization: Learning That Generalizes

AI 2025
SUMMER SCHOOL
ai.uni-jena.de

Niklas Maximilian Becker-Klöser

Faculty of Mathematics and Computer Sciences
Friedrich Schiller University Jena

Concept vs. Correlation: What We Mean by Generalization

Generalization is a model's performance on **unseen** data drawn from the **intended** target distribution. In modern, overparameterized regimes, model fit no longer maps cleanly to reliability: **test risk** — the expected loss on unseen target data — can follow a **double-descent** curve (falling, rising near interpolation, then falling again). Held-out **accuracy alone** does not **certify robustness under shift.** [2]

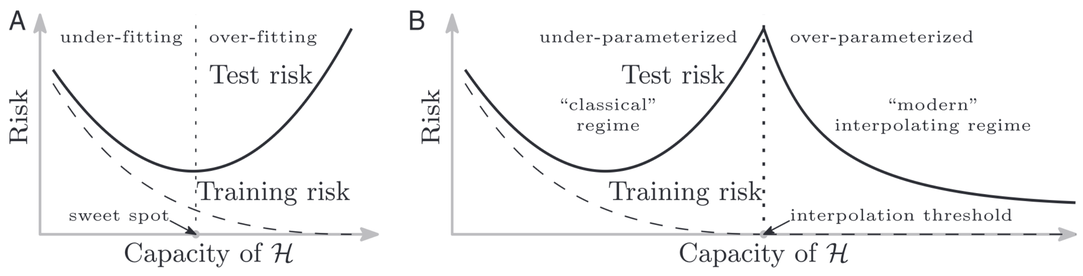

Fig. 1: Test risk vs. model capacity — classic U-curve and double-descent (fit can decouple from reliability). [2]

How Models Cheat: Memorization & Shortcuts

Deep nets can **perfectly fit random labels** and even noisy images — capacity enables memorization without concept learning. In practice, models often exploit *shortcuts* (textures, backgrounds, data-collection artifacts): non-causal features that inflate in-domain scores yet fail under mild shifts.

IID = *independent and identically distributed* evaluation; **OOD** = *out-of-distribution* evaluation used to measure robustness. [3, 5]

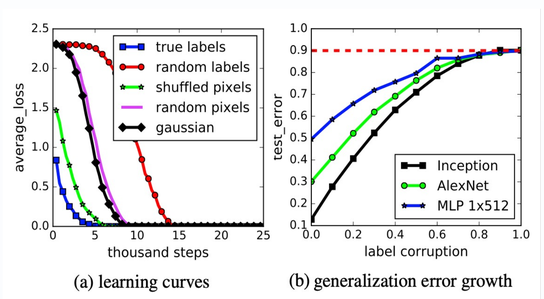

Fig. 2: Learning curves for true vs. permuted labels — train error → 0 while test error increases with label corruption (memorization exposed). [5]

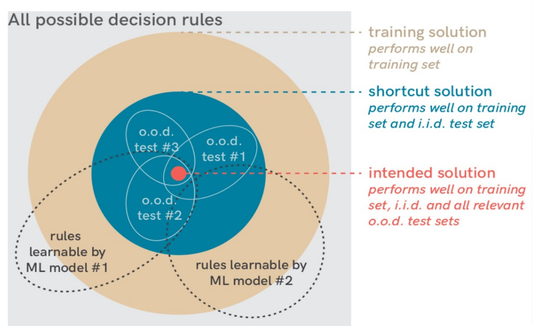

Fig. 3: Taxonomy of decision rules: all rules → training solutions → IID solutions; shortcuts contrasted with the intended (causal) rule. [3]

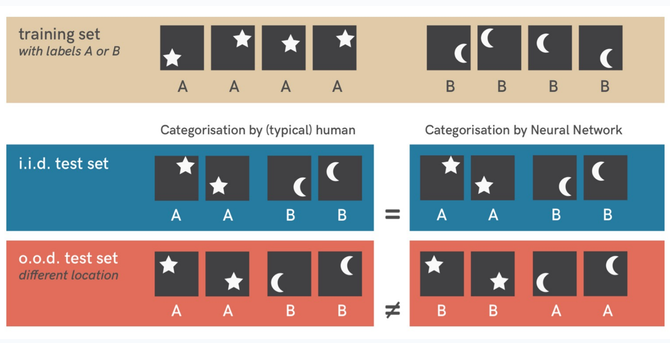

Fig. 4: Toy Shortcut Example [3]

A simple example of **shortcut learning** is shown in Figure 4. The model appears accurate in IID validation; however, OOD tests reveal that it prioritizes position over shape, leading to **systematic mislabeling**.

Diagnostics you can run:

(i) Compare training/test dynamics on *true vs. permuted labels*;

(ii) Visualize feature use (saliency/counterfactuals) to detect background or texture reliance. [3, 5]

Shift Happens: Evaluating Under Distribution Shift (OOD)

Held-out accuracy ≠ reliability. Add ***OOD*** tests that mirror deployment variation.

A. Corruptions & perturbations (ImageNet-C/P).

Standardized **corruptions** (weather/noise/blur, etc.) and small **perturbations** quantify stability.

Report: **mCE** (mean corruption error; lower is better) and flip/consistency rates. These stress tests reveal brittleness that clean accuracy can hide. [1]

Tutorial steps:

1. Run your baseline on **ImageNet-C** and compute **mCE** (per [1]);
2. Run **ImageNet-P** to measure stability under tiny input changes.
3. Compare both to clean accuracy and report them together. [1]

B. Real-world shifts (WILDS).

Choose datasets whose shifts match your domain (hospital, camera, geography, time).

Report the dataset-specific in-domain vs. OOD metrics using the **official protocol**. [6]

Tutorial steps:

1. Train and evaluate on the **in-domain** split.
2. Evaluate on OOD split(s).
3. Present the **gap** and discuss failure modes (e.g., subgroup, time drift). [6]

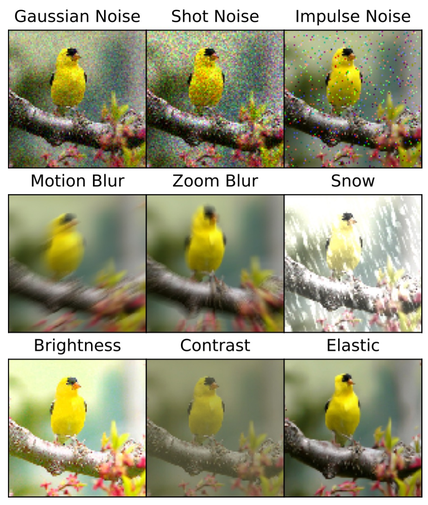

Fig. 5: ImageNet-C Corruption Families (Noise, Blur, Weather, Digital) [1]

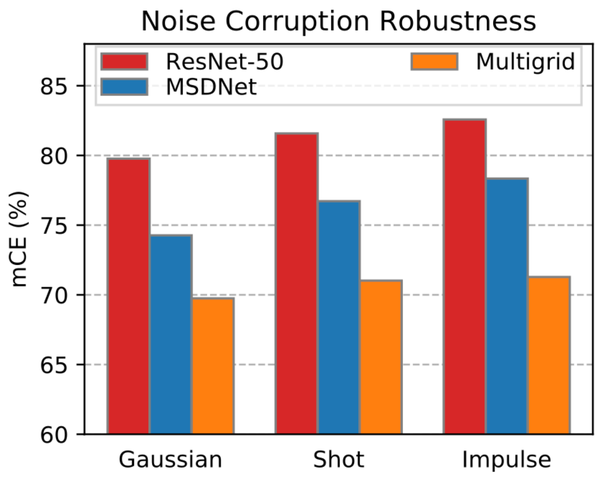

Fig. 6: Corruption robustness: mCE bars across architectures (lower is better). [1]

Figure 5 makes robustness tangible by showing some of the **ImageNet-C** corruption families (noise, blur, weather, digital), while Figure 6 distills the same idea **quantitatively** into **mCE** bar charts (lower is better), separating models that truly withstand disturbances from those that only shine on clean inputs. Together with the **WILDS Table 1** — real domains (hospitals, seasons, product categories, time) — this builds a coherent bridge from **controlled stressors** to **real-world shifts** — yielding a unified picture of **robustness** and **deployment-time generalization.** [1, 6]

Dataset	Metric	In-distribution type	In-distribution	Out-of-distribution
CAMELYON17-WILDS	Average accuracy	Fixed-train	93.2 (5.2)	70.3 (6.4)
GLOBALWHEAT-WILDS	Average domain accuracy	Fixed-test	64.8 (0.4)	48.4 (1.8)
CIVILCOMMENTS-WILDS	Worst-group accuracy	Average	92.2 (0.1)	56.0 (3.6)
FMOW-WILDS	Worst-region accuracy	Fixed-test	48.6 (0.9)	32.3 (1.3)
AMAZON-WILDS	10th percentile accuracy	Average	71.9 (0.1)	53.8 (0.8)

Tab. 1: IID → OOD drops across diverse domains — real-world shift consistently degrades performance. [6]

Same Score, Different Behavior: Underspecification & Stability

Entire pipelines can be **underspecified**: many predictors tie on validation accuracy yet behave **differently** on OOD data or subgroups due to seeds, initializations, or small design choices. This is not classical overfitting; it's an instability that standard splits don't reveal.

Evaluate **families** of models (multiple seeds/inits; slight architectural or data-processing variants) and **report dispersion** (mean ± std, or full distributions). Avoid conclusions from a single lucky run. [4]

Three-step recipe:

1. Fix data and training recipe; vary only **seeds/inits** (≥5 recommended).
2. For each trained model, record **validation** and **OOD** metrics.
3. Plot dispersion/correlation and **discuss divergence** even at similar validation accuracy. [4]

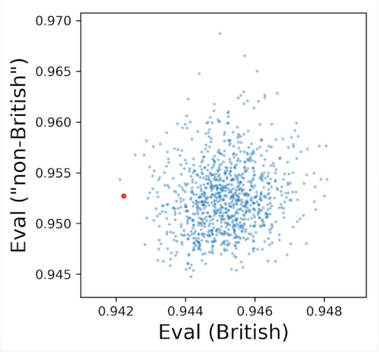

Fig. 7: Scatter of Eval (non-British) vs. (British) — similar IID scores, but divergent OOD behavior (low correlation). [4]

What Actually Helps: Practical Levers & Reporting

Test where you deploy: pair clean accuracy with *OOD* metrics (e.g., mCE; WILDS). [1, 6]

Reduce shortcuts: use data/augmentations and inductive biases that make spurious cues less useful; **validate under shift**, not only in-domain. [3]

Report stability: multiple seeds, dispersion, and ablations on data/training choices; treat stability as a first-class result. [4]

Capacity with context: interpret gains through the *double-descent* lens and OOD results — not only held-out accuracy. [2]

References

[1] D. Hendrycks and T. Dietterich, 'Benchmarking Neural Network Robustness to Common Corruptions and Perturbations', *ICLR*, 2019.

[2] M. Belkin, D. Hsu, S. Ma, and S. Mandal, 'Reconciling modern machine-learning practice and the classical bias–variance trade-off', *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 32, pp. 15849–15854, Aug. 2019, doi: 10.1073/pnas.1903070116.

[3] R. Geirhos *et al.*, 'Shortcut learning in deep neural networks', *Nat Mach Intell*, vol. 2, no. 11, pp. 665–673, Nov. 2020, doi: 10.1038/s42256-020-00257-z.

[4] A. D'Amour *et al.*, 'Underspecification Presents Challenges for Credibility in Modern Machine Learning', *Journal of Machine Learning Research*, vol. 23, no. 226, pp. 1–61, 2022.

[5] C. Zhang, S. Bengio, and M. Hardt, 'Understanding deep learning requires rethinking generalization', *ICLR*, 2017.

[6] P. W. Koh *et al.*, 'WILDS: A Benchmark of in-the-Wild Distribution Shifts', in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, July 2021, pp. 5637–5664. Accessed: Aug. 12, 2025. [Online]. Available: https://proceedings.mlr.press/v139/koh21a.html

Contact

niklas.maximilian.becker-kloeser@uni-jena.de