Online Resource 1: Experimental Results

Manuscript: Family Balancing and Hierarchical CNNs for Arabic Handwritten Symbols: A Controlled Five-Seed Evaluation

Author: Meshaal Mouawad
Department of Electrical and Computer Engineering, Mississippi State University
Contact: mm4922@msstate.edu
Experimental-record release: IJDAR-CNN-2026-10-28-R1

This release presents the same measured results as the main manuscript and connects them to the current record in Online Resource 3 (ESM_3.zip). Complete seed-level values are in results/full_210_metric_rows.csv; unrounded paired calculations are in results/PAIRED_HELDOUT_STATISTICS.json.

1 Design and units

Six configurations were trained with seeds 7, 17, 29, 43, and 61. Each selected model was evaluated on clean images and six fixed transformations of the same 13,360 test images (10,000 digits and 3,360 letters). The resulting 210 records are 30 model runs multiplied by seven conditions. Across-seed intervals describe five paired training replications. Corrupted conditions reuse the same ground-truth examples.

2 Clean-image outcomes

Table 1 reports mean ± sample standard deviation over five seeds. Accuracy and F1 are fractions. The two family F1 values are class averages from the complete 38-label confusion matrix; cross-family mistakes remain included.

Table 1: Clean test results; numeric values preserved from the main manuscript

Arm	Accuracy	Digit F1	Letter F1	Worst-family F1
B0: Flat / pooled	0.9835 ± 0.0032	0.9901 ± 0.0024	0.9637 ± 0.0109	0.9637 ± 0.0109
B1: Flat / balanced	0.9836 ± 0.0011	0.9896 ± 0.0014	0.9658 ± 0.0070	0.9658 ± 0.0070
B2: Hierarchical / pooled	0.9837 ± 0.0025	0.9902 ± 0.0023	0.9643 ± 0.0045	0.9643 ± 0.0045
B3: Hierarchical / balanced	0.9822 ± 0.0046	0.9880 ± 0.0036	0.9653 ± 0.0085	0.9653 ± 0.0085
P: Smooth-worst	0.9835 ± 0.0026	0.9898 ± 0.0021	0.9651 ± 0.0057	0.9651 ± 0.0057
C: Parameter-matched flat	0.9840 ± 0.0029	0.9903 ± 0.0024	0.9650 ± 0.0055	0.9650 ± 0.0055

3 Paired clean-image contrasts

Table 2 uses percentage points: each unrounded F1 difference and interval endpoint is multiplied by 100. A positive effect favors the first model. All five intervals cross zero. P−B3 is a supplementary *clean-image diagnostic*; the H3 endpoint is the corruption-retention comparison in Section 5, not this clean contrast.

Table 2: Paired means and descriptive 95% Student-t intervals over five seeds

Contrast	Change (pp)	95% interval (pp)
H1: family balancing (B1−B0)	+0.212	[-0.756, +1.179]
H2: hierarchy, pooled (B2−B0)	+0.054	[-1.370, +1.477]
H2: hierarchy, balanced (B3−B1)	-0.053	[-1.257, +1.152]
Supplementary clean diagnostic (P−B3)	-0.024	[-1.418, +1.370]
Parameter-count control (B2−C)	-0.077	[-0.969, +0.815]

4 Fixed-corruption outcomes

Gaussian noise uses intensity standard deviation $\sigma = 0.05$ or 0.10 on normalized pixels. Gaussian blur uses spatial standard deviation $\sigma = 0.8$ or 1.3 pixels. Contrast scales pixel distance from the midpoint 0.5 by $\alpha = 0.75$ or 0.50, then clips to $[0, 1]$. Thus, contrast reduction also raises a black background to 0.125 or 0.25, respectively. Table 3 reports condition means, not paired retention effects.

Table 3: Worst-family macro-F1, mean across five seeds

Condition	B0	B1	B2	B3	P	C
Clean	0.9637	0.9658	0.9643	0.9653	0.9651	0.9650
Noise $\sigma = 0.05$	0.0062	0.0100	0.0000	0.0134	0.0008	0.0053
Noise $\sigma = 0.10$	0.0000	0.0018	0.0000	0.0000	0.0000	0.0002
Blur $\sigma = 0.8$ px	0.9058	0.9347	0.8852	0.8850	0.8337	0.9268
Blur $\sigma = 1.3$ px	0.6011	0.7218	0.6788	0.5629	0.4928	0.7050
Contrast $\alpha = 0.75$	0.0359	0.0631	0.0883	0.0836	0.0262	0.0115
Contrast $\alpha = 0.50$	0.0004	0.0000	0.0006	0.0009	0.0008	0.0045

5 H3: equally weighted corruption retention

For each seed, the retention effect compares the corrupt-minus-clean change for P with the same change for B3, then averages with equal weight over the six corruptions. The mean effect is -0.03168 F1, with descriptive 95% interval $[-0.09548, +0.03212]$. This comparison is distinct from the clean P$-$B3 diagnostic.

The machine-readable path is `H3_smooth_worst/retention_differences/H3_prespecified_mean_across_6_corruptions`. Historical JSON key names are preserved. Their clean diagnostic does not change the designated H3 endpoint.

6 Evidence access and verification

Online Resource 3 contains all 30 prediction files, selected checkpoints, original epoch histories, unrounded results, the frozen implementation, and file identities. Its `README.md` provides commands for metric recalculation without image downloads or training. The new verification reconstructs confusion matrices directly from predictions, reproduces all 210 metric rows with zero absolute difference, checks the paired contrasts, and confirms 30 checkpoint-selection choices and 16 frozen code identities. Earlier selected-inference and partition checks are preserved separately as earlier audit evidence.

The original source data, predictions, checkpoints and experimental code were not modified for this supplement. This document updates labels and file navigation; it introduces no new training result.