ALR-CNN — Supervised Baseline Results

Model: ALR-CNN | Params: 515,329 | Input: 48 channels × 400 samples (200 ms) | Classes: 17 gestures (G13–G29) | Dataset: 87,614 windows | Val: 10% of training portion

Split	Train Size	Test Size	Val Size	Accuracy	Precision (W)	Recall (W)	F1 Score (W)	ROC-AUC (W)	Train Time (s)	Test Time (s)	Best?
90:10	**70,966**	**8,762**	**7,885**	**0.9879**	**0.9879**	**0.9879**	**0.9879**	**0.9999**	**4,077.2**	**1.04**	★
80:20	63,081	17,523	7,009	0.9847	0.9847	0.9847	0.9847	0.9998	1,088.7	1.60	
70:30	55,196	26,285	6,133	0.9793	0.9794	0.9793	0.9793	0.9998	798.7	2.11	
60:40	47,311	35,046	5,257	0.9767	0.9768	0.9767	0.9767	0.9997	833.6	2.65	
50:50	39,426	43,807	4,381	0.9550	0.9555	0.9550	0.9550	0.9992	584.6	3.28	
40:60	31,540	52,569	3,505	0.9615	0.9616	0.9615	0.9615	0.9993	499.5	4.07	
30:70	23,655	61,330	2,628	0.9390	0.9393	0.9390	0.9390	0.9985	615.8	4.87	
20:80	15,769	70,092	1,753	0.9308	0.9308	0.9308	0.9307	0.9981	305.2	6.11	
10:90	7,884	78,853	877	0.8875	0.8882	0.8875	0.8876	0.9951	187.0	5.81	

★ **Best supervised baseline (90:10 split, Accuracy / Weighted F1 = 0.9879, ROC-AUC = 0.9999). Use this as the SSL reference.**

Key Observations

• Performance degrades monotonically as training data decreases: from 0.9879 (90:10) down to 0.8875 (10:90), a drop of ~10 pp across the full range.
• The 40:60 split produced a slight uptick (0.9615) relative to 50:50 (0.9550), suggesting variance in stochastic training at mid-range data sizes.
• ROC-AUC remains very high across all splits (≥ 0.9951), indicating strong class separability even with limited training data.
• Training wall-clock time scales with dataset size; the 90:10 run took ~4,077 s due to the largest training set, while 10:90 ran in just 187 s.
• GFLOPs/window could not be computed in the Kaggle environment (reported as −1). Input size is fixed at 48 × 400 for all experiments.

Failure Mode Summary (Worst Confused Classes)

Across all splits the consistently lowest per-class accuracy is seen in G16, G19, G20, and G28, while G15 and G21 are the most reliably classified. Confusions are most frequent between gestures that share similar forearm posture (e.g. G16↔G19, G20↔G28). As training data shrinks (≤ 20%), additional classes (G14, G17, G23) also begin to degrade, indicating these are the most data-hungry gesture categories.

Best supervised baseline for SSL reference: 90:10 split (Weighted F1 = 0.9879)