Evaluation of 3D Convolutional Neural Networks on Isolated Brazilian Sign Language Recognition

Lorenzo C. Lazzarotto*, Marcelo M. Delucis,
Pedro T. Barcelos, Rodrigo C. Barros, Lucas S. Kupssinskü

[1]Technology School – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brazil

l.lazzarotto@edu.pucrs.br

Abstract. *Isolated Sign Language Recognition plays a crucial role in advancing technologies that enhance accessibility for deaf individuals. In this work we investigate the performance of several 3D Convolutional Neural Networks on the task of ISLR in LIBRAS. Models were trained on the MINDS dataset and evaluated on an out-of-domain dataset, MALTA-LIBRAS. Experimental results indicate that while all models achieve high metrics on the training and validation data, they exhibit clear overfitting and fail to generalize effectively to unseen samples. These results highlight the challenges posed by limited data and data variability in ISLR for LIBRAS.*

1. Introduction

The World Health Organization estimates that roughly 430 million people live with disabling hearing loss, and the World Federation of the Deaf reports that about 70 million people are members of signing communities [World Health Organization 2021, World Federation of the Deaf 2024]. Sign Languages (SLs) are natural, full-fledged languages with their own phonology, morphology, and syntax; however, the lack of accessible tools still hinders communication between signers and non-signers.

In Brazil, it is estimated that up to 10 million people has some level of hearing loss from which only 270 thousand utilize Brazilian Sign Language (LIBRAS) as a mean of communication, making the development of computational tools useful for access to services, education, and work [Paula Pfeifer Moreira 2025]. Despite this societal need, research on automatic LIBRAS recognition remains constrained by small datasets and limited capture diversity, which often leads to high in-domain accuracy but poor generalization to new signers, environments and signs [Delucis 2025].

For our experiments, we adopt a cross-dataset protocol: models are trained and validated on the MINDS [Rezende et al. 2021] dataset and evaluated out-of-domain (OOD) on MALTA-LIBRAS [Delucis 2025] dataset, a curated collection that matches the same 20 glosses but introduces different signers, cameras, and backgrounds. This design directly probes robustness rather than just fitting the homogeneous conditions of MINDS.

Ultimately, we benchmark the I3D-R50 [Carreira and Zisserman 2017], X3D-S [Feichtenhofer 2020] and Slow-R50 [Feichtenhofer et al. 2019] under a training pipeline and report macro top-1 accuracy, precision, recall, and F1. Our contributions are: (i) a baseline for LIBRAS ISLR with CNN backbones; (ii) a cross-dataset evaluation from MINDS to MALTA-LIBRAS quantifying generalization to unseen conditions.

*Corresponding author.

2. Methodology

2.1. Task and Evaluation Protocol

We formulate LIBRAS isolated sign recognition (ISLR) as a 20-way video classification problem. Given the scarcity of labeled LIBRAS video corpora, we train and validate on MINDS [Rezende et al. 2021], which is , to our knowledge, the most recent and field-relevant LIBRAS ISLR dataset. While MINDS enables controlled benchmarking, it exhibits limited capture variability (uniform background, lighting, and camera) and, compared with widely used ASL corpora, such as the Word-Level American Sign Language (WLASL) dataset [Li et al. 2020], is smaller, with fewer classes and samples per class.

2.1.1. Datasets

MINDS This dataset consists of 1.158 RGB clips covering 20 signs performed by 12 signers, captured under controlled background, lighting, and camera settings. We adopt a stratified $75/25$ train-validation split per class.

MALTA-LIBRAS To introduce variability absent from MINDS, the MALTA-LIBRAS dataset aggregates videos from six publicly available sources. It matches the same 20 signs but with different signers, resolutions, and recording conditions. The intersection used for testing contains 132 clips, which we use exclusively as an OOD test set.

2.1.2. Models

X3D X3D [Feichtenhofer 2020] is a family of video recognition models that progressively expand a baseline MobileNet-style [Howard et al. 2017] 2D image classification architecture. The expansion occurs along multiple network axes: space, time, width, and depth. We evaluate three configurations (XS, S, M) that share the same parameter count (3.79M) but differ in total computational cost. The Floating Point Operations Per Second (FLOPs) for the XS, S and M configurations are, respectively, 0.91G, 2.96G and 6.72G.

Slow R50 The Slow R50 model is the "slow-only" pathway of the SlowFast family, a 3D variant of ResNet-50 for video recognition [Feichtenhofer et al. 2019]. It operates on sparsely sampled frames (low frame rate), prioritizing high-fidelity spatial semantics while accumulating long-range temporal context through temporal convolutions. By inflating the 2D ResNet-50 kernels into the temporal dimension, it becomes a spatiotemporal feature extractor with approximately 32.45M parameters.

I3D The I3D-R50 model is a ResNet-50–based instantiation of Inflated 3D ConvNets [Carreira and Zisserman 2017]. It "inflates" 2D convolution and pooling kernels into 3D, adding a temporal extent, typically initializing from ImageNet weights via kernel inflation. This yields a model that learns spatiotemporal features end-to-end for video recognition and contains approximately 28.04M parameters.

2.1.3. Experimental Design and Setup

Videos were resized to 224×224, rescaled to $[0, 1]$, and normalized with Kinetics-400 statistics [Kay et al. 2017]. Clips are then uniformly subsampled to 64 frames. For our optimizer we chose the AdamW [Loshchilov and Hutter 2019] with learning rate of
1×10^{-5} for 200 epochs with an early stop of 10 epochs on the validation loss. We select the best checkpoint by MINDS validation loss and test it on the MALTA-LIBRAS dataset.

All models are trained on MINDS and the checkpoint with the lowest validation loss is selected to perform a single OOD evaluation on the MALTA-LIBRAS dataset. Our primary objective is to, for each model, isolate the contribution of Kinetics-400 pretraining relative to training from random initialization.

In our work we report the pretrained baselines for all models. Our setup is comprised of an Nvidia Geforce GTX 1080Ti and an Nvidia RTX A6000. Reported metrics are macro-averaged top-1 accuracy, precision, recall, and F1-score.

3. Results & Discussions

Table 1. Validation results in the MINDS dataset and test results on the MALTA-LIBRAS dataset.

	MINDS Validation Set				MALTA Test Dataset			
Model	Accuracy	Recall	Precision	F1 Score	Accuracy	Recall	Precision	F1 Score
X3D XS	0.993	0.993	0.994	0.993	0.144	0.150	0.133	0.101
X3D S	0.996	0.996	0.997	0.996	0.379	0.367	0.524	0.347
X3D M	**1**	**1**	**1**	**1**	0.318	0.307	0.443	0.291
I3D R50	0.986	0.986	0.989	0.986	**0.455**	**0.439**	**0.619**	**0.428**
Slow R50	0.983	0.983	0.985	0.983	0.386	0.371	0.605	0.371

As shown in Table 1, all models achieve near-perfect metrics on the MINDS validation split; X3D-M reaches 1 for all metrics, with the others only a few points behind. These results indicate that in-domain validation is a weak indicator of robustness.

On the OOD MALTA-LIBRAS test (with random chance $= 0.05$), metrics diverges: I3D-R50 attains the best overall metrics, followed by Slow R50 and X3D-S. Although X3D-S matches Slow R50 in accuracy, its precision is lower. X3D-M generalizes worse than X3D-S, and X3D-XS performs worse. Thus, near-perfect MINDS validation does not carry over to modest shifts in signer, camera, and background, making cross-dataset generalization the central challenge.

The OOD metrics reveal clear trade-offs: I3D-R50 yields the highest F1 (best balance of precision and recall), whereas Slow R50 is more conservative (higher precision, lower recall). Within X3D, moving from XS to S improves OOD, but M adds capacity without gains, consistent with overfitting to MINDS' controlled protocol and limited signer variability. The gap between MINDS and MALTA-LIBRAS shows that capacity and temporal footprint alone do not ensure generalization.

The primary bottleneck is MINDS severe controlled conditions. Under our constraints we report pretrained RGB models without augmentation or two-stream fusion; OOD results remain above chance (I3D-R50 accuracy $= 0.455$), indicating partial transfer from generic video pretraining to LIBRAS ISLR.

4. Final Considerations

In this work we evaluated 3D CNNs for LIBRAS ISLR under the MINDS validation split and MALTA-LIBRAS test split protocol. Consistent with prior studies, models fit the homogeneous MINDS split but metrics decline on MALTA-LIBRAS, underscoring the generalization challenge [Delucis 2025]. Our results with off-the-shelf action-recognition pretraining align with recent transformer baselines: pretraining is necessary to rise metrics to above chance on MALTA-LIBRAS, highlighting its value under data scarcity. Relative to those transformer baselines, our models achieve higher OOD accuracy, indicating that backbone choice, along with pretraining, affects LIBRAS ISLR model performance.

Nevertheless, our findings support prior work: limited variability in current LIBRAS corpora induces overfitting, reinforcing the need for more data. Future work will prioritize (i) systematic video augmentations, (ii) signer-disjoint protocols and larger, more diverse LIBRAS datasets, and (iii) alternative modalities such as two-stream RGB+optical flow, which prior studies indicate can mitigate overfitting.

Acknowledgments This study was financed by Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil - (CNPq) - Grant Number: 443072/2024-8.

References

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Delucis, M. M. (2025). Isolated Sign Language Recognition in LIBRAS. Master's thesis, Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv*.

Li, D., Opazo, C., Yu, X., and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1448–1458, Los Alamitos, CA, USA. IEEE Computer Society.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*, page 18. OpenReview.

Paula Pfeifer Moreira (2025). Quantos surdos usam libras no brasil em 2025. Accessed in: August 2025.

Rezende, T. M., Almeida, S. G. M., and Guimarães, F. G. (2021). Development and validation of a brazilian sign language database for human gesture recognition. *Neural Computing and Applications*, 33(16):10449–10467.

World Federation of the Deaf (2024). Our work. Accessed in: Nov 2023.

World Health Organization (2021). *World Report on Hearing*. World Health Organization.