



POLITECNICO
MILANO 1863

**SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE**

EXECUTIVE SUMMARY OF THE THESIS

AI for inclusivity: a deep learning approach for Italian Sign Language Recognition

LAUREA MAGISTRALE IN COMPUTER SCIENCE AND ENGINEERING

Author: PAOLO BIOLGHINI

Advisor: PROF. MARISTELLA MATERA

Co-advisor: LUDOVICA PIRO, MARCO LA CAMERA

Academic year: 2024-2025

1. Introduction

The inability of many people to understand sign language represents a significant communicative barrier for the Deaf community. Research on automatic sign language recognition aims to overcome this barrier, fostering more inclusive and effective communication. This study proposes a vision-based approach that relies on the extraction of skeletal keypoints, from which 42 joint angles are computed and used as input for the analyzed Machine Learning and Deep Learning models. In particular, the performance of 1D convolutional models, bidirectional LSTM architectures enhanced with attention mechanisms, and Transformer-based architectures has been evaluated. The work addresses three main tasks: isolated sign classification, out-of-distribution sign recognition, and continuous sign sequence classification. Experimental analyses were conducted on two private datasets containing 46 and 72 Italian Sign Language (LIS) signs, respectively.

2. Related Works

Automatic Sign Language Recognition (ASLR) has advanced significantly, driven by advances

in machine learning and computer vision. Approaches can be divided by task type (isolated vs continuous signs recognition) and by the method used for data acquisition (sensor-based or vision-based). Isolated recognition deals with single signs per input, simplifying segmentation, while continuous recognition involves sequences of multiple signs.

Sensor-based approaches employ wearable devices to capture hand kinematics with high precision; however, they are limited in capturing facial expressions and whole-body posture, and their intrusiveness hinders adoption within the Deaf community. Vision-based methods leverage camera systems to enable the extraction of both manual and non-manual components while reducing costs and offering a more accessible modality for data acquisition. However, their performance can be adversely affected by challenging lighting conditions, background biases, and signer-related variability. Such approaches can be categorized into three main paradigms: (i) frame-based methods, where the raw image frames are directly processed by the network; (ii) skeleton-based methods, in which the classification model operates on 2D or 3D keypoints representing the signer's body; and (iii) hybrid

approaches, which integrate both modalities to exploit their complementary strengths.

Classification models differ according to the type of input data. For sensor-based inputs, traditional machine learning techniques have been widely employed, with more recent approaches increasingly adopting deep learning methods. In contrast, vision-based data primarily rely on deep learning architectures, including CNNs, LSTMs, Transformers, GCNs, as well as multimodal fusion strategies for hybrid representations.

For instance, Zhou et al. [4] explore a sensor-based approach, evaluating multiple classical machine learning classifiers alongside an LSTM with attention mechanisms. Similarly, Kumari and Anand [1] investigate a vision-based method where a lightweight CNN based on MobileNetV2 is integrated with an attention-enhanced LSTM. Shin et al. [3] propose a simplified pipeline in which hand keypoints are extracted via Mediapipe [2], geometric features such as joint distances and angles are computed, combined with representations obtained from GoogleNet, and subsequently classified using an SVM.

A key challenge in ASLR is dataset availability. Sensor-based datasets are small and device-specific, while vision-based datasets are limited in size and diversity. Resources for Italian Sign Language (LIS) remain scarce, motivating the need for larger, more comprehensive datasets.

This study represents one of the first works in the Italian context employing a dataset with over 70 unique classes and more than 100 samples per sign, focusing on everyday vocabulary rather than domain-specific lexicons.

3. Methods

The work presented in this study is organized around three main tasks: (i) isolated sign recognition, (ii) out-of-distribution (OOD) detection, and (iii) continuous sign recognition. Each task is designed to address a different aspect of automatic sign language understanding, leveraging a unified feature representation extracted from raw video data.

3.1. Dataset

Two private datasets were collected in collaboration with Cludia Research, comprising 46 and 72 signs respectively, with a focus on ev-

eryday vocabulary rather than domain-specific terminology. Each video was recorded at 30 fps and processed using MediaPipe [2] to extract 2D skeletal keypoints from each frame. From these keypoints, 42 angular features were computed, capturing both manual and non-manual components of the signs. All features were normalized to the range $[-1, 1]$ to ensure numerical stability during training.

For both datasets, the samples were stratified into training (80%), validation (10%), and test (10%) sets, preserving a balanced class distribution across splits.

3.2. Isolated Sign Classification

Classical Machine Learning algorithms, including Logistic Regression, Support Vector Machines, Decision Trees, and Random Forests, were initially evaluated to establish performance baselines and assess non-neural approaches. No extensive hyperparameter tuning was conducted, as the objective was solely to obtain reference results.

Subsequently, a baseline *Multi-Layer Perceptron (MLP)* with three fully connected layers, combined with Dropout and Batch Normalization for regularization, was implemented. This model demonstrated the capability of neural networks to learn meaningful representations even in relatively simple configurations, thereby motivating the exploration of more advanced architectures. To further improve performance, three main families of models were investigated:

- **1D Convolutional Neural Networks (1D-CNNs)**: Designed to capture local temporal patterns efficiently. At the beginning of the network, multiple convolutional blocks are positioned, each consisting of one Conv1D layer with varying kernel sizes, Batch Normalization, ReLU activation, and Max Pooling. After flattening, fully connected layers with Dropout performed the final classification. The best configuration employed four convolutional blocks (kernel sizes 7, 5, 3, 3), two dense layers (512, 128 units), and Dropout rates of 0.4 and 0.3.
- **Long Short-Term Memory Networks (LSTMs)**: Given the sequential nature of sign language, both unidirectional and bidirectional LSTMs have been evaluated. Attention mechanisms were imple-

mented to assign appropriate weights to the information from different frames. The most effective model combined a Bi-LSTM layer (hidden size 128) with an attention module and fully connected layers with dropout.

- **Transformer-based Models:** Finally, Transformer encoders were explored to model long-range dependencies via self-attention without recurrence. Inputs were projected into a latent space with positional embeddings, followed by multi-head self-attention layers and feed-forward networks with residual connections and layer normalization. The optimal architecture used a latent dimension of 256, three encoder layers with eight attention heads, and a dropout rate of 0.2.

In addition to the standard *Categorical Cross-Entropy* (with class weights for imbalance handling), the use of *Contrastive Loss* and *Triplet Loss* was also explored, aiming to promote structured latent representations through metric learning. Hybrid losses combining Cross-Entropy with metric-based objectives were also tested to integrate supervised classification and semantic embedding learning.

All models were trained for a maximum of 300 epochs with *early stopping* (patience 10–20) and the *Adam* optimizer. Learning rate scheduling employed *ReduceLROnPlateau* or *Cosine Annealing* strategies. Hyperparameters, including learning rate, hidden dimensions, dropout rates, number of layers, and batch size, were tuned via *Bayesian optimization using Optuna*, ensuring a fair and efficient comparison across models.

3.3. Out-of-Distribution Detection

The goal of this second task is to determine whether a given input belongs to the set of known classes or constitutes an out-of-distribution sample, thereby enhancing the robustness of a classification system when encountering previously unseen signs. This capability is particularly important due to the limited vocabulary currently supported and the intrinsic variability of Italian Sign Language (LIS). By explicitly identifying unknown signs rather than forcing an incorrect classification, the system can maintain reliability and accuracy in open-set scenarios.

3.3.1 Embedding-Based Anomaly Detection

For this first method a two-stage pipeline was designed: a 1D-CNN first learns discriminative embeddings using Contrastive or Triplet Loss to encourage intra-class compactness and inter-class separation. Classical anomaly detection methods (K-means, DBSCAN, Isolation Forest) were then applied to identify outliers in the embedding space. Results highlighted the limitations of these techniques, motivating more advanced approaches.

3.3.2 Logit-Based Detection

Further analysis was conducted on the logits of the classification network trained with a combined Cross-Entropy and metric learning loss. Three uncertainty measures were considered: maximum probability, softmax entropy score, and logit energy score. Each metric is a threshold to decide whether a sample is out-of-distribution. Temperature scaling was applied to adjust confidence calibration.

3.3.3 Teacher-Student Framework

Finally, a Teacher-Student approach was employed, where a high-capacity Teacher guides a lighter Student via a Distillation Loss. The Student is expected to mimic the Teacher on in-distribution data but to fail on anomalies, enabling reliable OOD detection.

3.4. Multi-Sign Classification

To enable real-time translation of continuous Italian Sign Language (LIS) sequences, it was necessary to transition from single-sign classification to multi-sign recognition, where sign boundaries are unknown, reflecting real-world conditions. Each video is represented as a temporal stream of 42 features extracted using MediaPipe [2].

Three main approaches were investigated:

3.4.1 Sliding-Window Baseline

As a naive baseline, single-sign classifiers were applied to overlapping temporal windows extracted from the continuous stream. Window length and stride were optimized using Optuna, with zero-padding applied when neces-

sary. Post-processing merged consecutive identical predictions to improve output readability.

3.4.2 Reduced-Window Classification with Major Voting

To reduce train–test discrepancies, smaller windows were adopted both in training and inference. During inference, each window was split into multiple sub-sequences, whose predictions were aggregated using a *major voting* scheme. Models included CNN and LSTM architectures trained with combined triplet and cross-entropy losses to ensure both separability and classification accuracy.

3.4.3 End-to-End LSTM Models

Finally, end-to-end Long Short-Term Memory (LSTM) networks were explored to directly map entire sequences to sign labels without explicit segmentation. A synthetic dataset of continuous sign sequences was generated by concatenating isolated signs with transition frames to simulate temporal continuity. Variants included attention mechanisms and different architectural configurations to assess trade-offs between accuracy and model complexity. While promising, this approach required careful regularization to mitigate overfitting due to limited annotated data.

4. Result

4.1. Isolated Sign Classification Results

To evaluate model performance on isolated sign classification, accuracy and F1-score were employed as the primary metrics. The reported results refer exclusively to the test sets.

Initial experiments employed classical Machine Learning algorithms such as Logistic Regression, SVM, Decision Trees, and Random Forests. Random Forests achieved the best results, with 92.4% accuracy on the 46-class dataset and 82.7% on the 72-class one (Table 1). A simple Multilayer Perceptron (MLP) with two hidden layers (512–128 units) showed a slight improvement over simpler models but remained below Random Forest performance.

Next, deep learning approaches, including 1D-CNNs, LSTMs, and Transformers, were evaluated, with hyperparameters optimized using Op-

Table 1: Performance of the models on the two datasets (46 and 72 classes) in terms of Accuracy, Weighted F1.

Modelli ML	46 classi		72 classi	
	Acc.	F1	Acc.	F1
Log. Regression	84.5	84.6	66.0	65.9
SVM	87.7	87.8	71.4	71.6
Decision Trees	75.1	74.9	50.5	47.9
Random Forest	92.4	92.6	82.7	82.4
MLP	87.1	87.0	73.2	72.9

tuna. All models outperformed classical methods, with LSTMs achieving the best results: 95.4% on the 46-class dataset and 91.1% on the 72-class dataset (Table 2). Analysis of the results revealed a few ambiguities between visually similar signs, such as digits and letters sharing hand shapes.

Experiments investigated the effect of combining categorical cross-entropy with metric learning losses—Siamese and Triplet Loss—to improve class separability in the embedding space. Results (Table 2) showed consistent accuracy gains, with Triplet Loss yielding the best performance (up to 96.95% on 46 classes, 92.9% on 72 classes). Among all tested models, LSTMs consistently achieved the highest accuracy, closely followed by Transformers and 1D-CNNs.

Table 2: Performance of Deep Learning models with different loss functions on the two datasets (46 and 72 classes), in terms of Accuracy and Weighted F1.

Modello + Loss	46 classi		72 classi	
	Acc.	F1	Acc.	F1
1D-CNN + CE	94.4	94.6	86.2	85.4
1D-CNN + CE+Siamese	95.47	95.7	89.1	88.5
1D-CNN + CE+Triplet	96.0	96.2	90.37	89.8
LSTM + CE	95.4	95.6	91.1	90.9
LSTM + CE+Siamese	95.5	95.7	91.6	91.3
LSTM + CE+Triplet	96.95	97.1	92.9	92.6
Transf. + CE	95.2	95.2	90.9	90.7
Transf. + CE+Siamese	96.40	96.59	92.7	92.4
Transf. + CE+Triplet	96.4	96.6	92.0	91.6

4.2. Out Of Distribution Detection

For this second task, the experiments were conducted on the 46-sign dataset by splitting it into 40 in-distribution (ID) and 6 OOD classes, randomly selected to simulate previously unseen signs.

A first baseline relied on CNN-extracted embeddings (trained with categorical cross-entropy + triplet loss) combined with classical anomaly detection algorithms: Isolation Forest, K-means, and DBSCAN. Results show moderate performance, with Isolation Forest achieving the best OOD accuracy (67.0%), followed by K-means (60.5%) and DBSCAN (57.0%). However, overlapping embedding distributions highlight the limited separability of OOD samples under this approach, motivating the need for more advanced strategies.

Next, methods leveraging predictive uncertainty were explored, including maximum probability, softmax entropy score and logit energy score, combined with temperature scaling and threshold optimization. Across all architectures, LSTM consistently achieved the highest AUROC (up to 0.8705 with entropy) and OOD accuracy (76.5%), while Transformers performed moderately (AUROC 0.80) and 1D CNNs showed the lowest scores (AUROC 0.69–0.75). Among criteria, Entropy proved the most robust, confirming that uncertainty-aware metrics enhance OOD detection compared to simple confidence scores.

Finally, in the Teacher–Student framework, a lighter student network was trained via knowledge distillation from a high-performing teacher. The teacher networks corresponded to the top-performing isolated sign classification models of each class of architecture. Divergences between teacher and student predictions were used for OOD detection. Distillation consistently improved student accuracy (e.g., Transformers: 94.0% vs. 91.5% without distillation) and yielded competitive OOD detection (AUROC = 0.7950, TPR = 85%). Nevertheless, overall OOD accuracy remained below 80%, indicating performance limitations even with this paradigm.

4.3. Multi-Sign Classification

The third experimental phase addressed the most challenging task: translating continuous

sequences of Italian Sign Language (LIS) signs into textual units. A dedicated test set was created, including both alphabet sequences and sign sequences from the training set, enabling robust evaluation across different complexity levels. Performance was assessed using standard metrics such as *Precision*, *Recall*, *F1-score* and *Word Error Rate* (WER) to capture transcription accuracy.

The first approach extended isolated sign classification models (CNN, LSTM, Transformer) to continuous sequences using a fixed *sliding window* segmentation (30 frames). Models were trained with a joint *triplet loss* and *cross-entropy* objective. Results (Table 3) reveal severe performance degradation (F1-score < 21%, WER > 2.0), highlighting difficulties in both sign recognition and boundary detection. This confirmed the method’s role as a baseline for more advanced solutions.

Table 3: Baseline results with fixed windows.

Model	Precision	Recall	F1-score	WER
LSTM	25.5	16.6	20.2	2.21
CNN	20.3	14.5	16.9	2.38
Transformer	24.9	11.0	15.3	2.17

To better align training and inference conditions, sequences were segmented using shorter windows (15 frames) and predictions combined via *major voting*. This reduced temporal mismatch significantly improved performance (Table 4), with the LSTM model achieving the best results (F1-score = 76.0, WER = 0.71).

Table 4: Results with reduced windows and major voting.

Model	Precision	Recall	F1-score	WER
CNN	66.1	58.2	61.9	1.08
LSTM	78.2	74.1	76.0	0.71

Finally, end-to-end BI-LSTM models with attention were trained on synthetic sequences simulating continuous signing. These models achieved near-perfect performance on synthetic data (F1 = 97.1) but degraded on real videos (F1 = 69.8) due to domain mismatch (Table 5).

Overall, results indicate that *window-based* methods benefit from reduced window sizes and voting schemes, while *end-to-end* models show

Table 5: End-to-end BI-LSTM results on synthetic vs. real data.

Dataset	Precision	Recall	F1-score
Synthetic	97.6	96.8	97.1
Real	70.3	69.4	69.8

strong potential but require larger, domain-specific datasets for robust real-world deployment.

5. Conclusions

This study presented a comprehensive investigation into automatic Italian Sign Language (LIS) recognition, addressing isolated sign classification, out-of-distribution (OOD) detection, and continuous sign recognition within a unified framework. Our experiments highlighted several key findings.

First, deep learning models substantially outperformed classical machine learning approaches, with Bi-LSTMs consistently achieving the highest accuracy across isolated sign classification tasks. Moreover, integrating metric learning objectives, such as Triplet Loss, enhanced latent space separability, further improving classification performance.

Second, OOD detection remains challenging. While entropy-based uncertainty measures and teacher–student distillation frameworks improved robustness compared to classical embedding-based methods, overall accuracy on unseen signs remained below 80%, underscoring the need for more sophisticated approaches or larger datasets to handle open-set conditions effectively.

Finally, for continuous sign recognition, window-based models with reduced window sizes and majority voting yielded the best trade-off between segmentation accuracy and recognition performance, whereas end-to-end architectures demonstrated promising potential but suffered from domain mismatch when applied to real-world data.

Overall, our findings suggest that robust LIS translation systems require both advanced sequence modeling techniques and expanded, domain-specific datasets. Future work will focus on scaling data collection, exploring advanced skeleton-based architectures such as Graph Con-

volutional Networks, multimodal feature fusion with RGB data, and grammar-aware post-processing to improve syntactic consistency.

References

- [1] Diksha Kumari and Radhey Anand. Isolated video-based sign language recognition using a hybrid cnn-lstm framework based on attention mechanism. *Electronics*, 13:1229, 03 2024.
- [2] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019.
- [3] Jungpil Shin, Md Al Mehedi Hasan, Abu Saleh Musa Miah, Kota Suzuki, and Koki Hirooka. Japanese sign language recognition by combining joint skeleton-based hand-crafted and pixel-based deep learning features with machine learning classification. *Comput. Model. Eng. Sci*, 139(3):2605–2625, 2024.
- [4] Zhihao Zhou, Kyle Chen, Xiaoshi Li, Songlin Zhang, Yufen Wu, Yihao Zhou, Keyu Meng, Chenchen Sun, Qiang He, Wenjing Fan, Endong Fan, Zhiwei Lin, Xulong Tan, Weili Deng, Jin Yang, and Jun Chen. Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. *Nature Electronics*, 3(9):571–578, 2020.