Journal on Interactive Systems, 2026, 17:1
ISSN: 2763-7719 • doi: 10.5753/jis.2026.6369

Research Paper

Use of Neural Networks to Classify Real Time Signs From Brazilian Sign Language

Luciano Alves Cardona Junior [Federal University of Santa Maria | *cardonixjr@gmail.com*]
Anselmo Rafael Cukla [Federal University of Santa Maria | *anselmo.cukla@ufsm.br*]
Solon Bevilacqua [Federal University of Goiás | *solon@ufg.br*]
Daniel Pinheiro Bernardon [Federal University of Santa Maria | *dpbernardon@ufsm.br*]
Daniel Fernando Tello Gamarra [Federal University of Santa Maria | *daniel.gamarra@ufsm.br*]

Department of Electric Energy Processing (DPEE), Universidade Federal de Santa Maria, Av. Roraima nº 1000, Room 1201 A, Annex C of the Technology Center, Camobi, Santa Maria, RS, 97105-900, Brazil.

Abstract. This study presents the use of convolutional neural networks aimed at detecting and simultaneously translating continuous signs of Brazilian Sign Language (LIBRAS) in videos captured by common computer cameras. The objective of this work was to propose a methodology that enabled the classification of signs performed by people of different genders, body types, and skin tones, with minimal interference from the video background while using cameras of medium or low quality, creating an accessible approach for various applications. For this purpose, the MediaPipe algorithm was used for video data extraction, the FastDTW algorithm was employed for data standardization, convolutional neural networks were implemented using the TensorFlow and Keras libraries for sign recognition, and a custom dataset was created to train the network. All these tools were integrated using the Python programming language to produce a model for real-time classification of continuous LIBRAS signals.

Keywords: Neural Network, Sign Classification, Brazilian Sign Language, Real-Time Detection.

Edited by: Saul Delabrida | **Received:** 18 August 2025 • **Accepted:** 02 July 2026 • **Published:** 12 July 2026

1 Introduction.

Brazil has one of the most advanced sets of laws for accessibility, and the approval of the law ”Lei Brasileira da Inclusão” (Brasil [2015]) and Brazilian Decree No. 5.626/2005 (Brasil [2005]) brought resolutions that, according to Rumjanek and da Silva [2019] and Castro *et al*. [2011], support the diffusion of Brazilian Sign Language and create opportunities for technologies aimed at accessibility and inclusion, presenting several improvements to allow better opportunities for people with disabilities, including the deaf community. However, as these authors stated, there are several difficulties in implementing these laws.

According to the Instituto Brasileiro de Geografia e Estatística [2019] national health survey reviewed by Stopa *et al*. [2020], only 1.8% of people with moderate hearing impairments know how to use Brazilian Sign Language (LIBRAS), and even among those with total hearing loss, this percentage is only 35.8%. So, even if authors like Córdova [2018] and Courtin [2000] claim that LIBRAS is an official (native) brazilian language system, communication between deaf and speaking people presents some failures when speakers do not have enough knowledge of the sign languages, as evidenced by Hommes *et al*. [2018]. These data highlight the poor education of the population in one of Brazil's official languages, even among a group that needs this tool to communicate daily, exposing one of the barriers people with hearing disabilities face in social communication, especially when they need to establish this social contact with people who do not know LIBRAS, as denoted by Abdallah and Fayyoumi [2016].

Furthermore, Antunes *et al*. [2011] highlights that LIBRAS is not a simple spelling of spoken language, and has notable differences from written and spelled Portuguese in grammar, morphology, syntax, and semantics, so deaf people struggle to use Portuguese to communicate their needs, although they might know some words in Portuguese. Rocha *et al*. [2020] explains that this language barrier increases because the deaf generally do not master written language, and only a few people with hearing can communicate using sign language.

Therefore, as it advocated by Palavissini *et al*. [2021], De Oliveira Feliciano *et al*. [2023] and Bastos *et al*. [2015], is vital to propose new perspectives of human-machine interaction focused on the direct interpretation of LIBRAS signs, applied to existing technologies of communication, entertainment, education, and information in various social contexts, including domestic and educational.

Given this scenario, Trotta *et al*. [2022], Silva *et al*. [2023] and Dias *et al*. [2024] points out that artificial neural network-based video detection and interpretation algorithms emerge as a possible solution to these challenges . Such tools offer solutions that employ convolutional neural networks to recognize LIBRAS signs performed by a user in front of a camera or sensor and, based on these data, make relevant decisions for the device.

So, the main contributions of this work are: 1) the proposal of a methodology for detecting and classifying continuous signs in Brazilian Sign Language; 2) the appliance of the neural network model in real-time scenarios, without depending on the environment, the characteristics of the people who perform the sign or the quality of the camera that recorded it; 3) the detection of 20 different LIBRAS signs in real-time using the proposed architecture; 4) the availability of trained neural network models, useful open-source tools for dataset construction, data processing and network train-

ing; 5) a proof of concept for a pipeline for the application of open-source techniques that could be used in future works to improve communication and interaction of deaf individuals, hearing-impaired people, or native LIBRAS speakers.

The article will be divided into six sections. After a succinct introduction, there is a related works section. The third section then presents the theoretical background that guides this study, the fourth section will discuss the methodology, the fifth section shows the results obtained, and finally, the last section discusses the main conclusions achieved in the work.

2 Related Works.

This study starts from the assumptions made by Rumjanek and da Silva [2019] and Castro *et al*. [2011], who state that although Brazilian legislation promotes accessibility and inclusion of individuals with hearing impairments, it fails to ensure their practical integration into society, including the lack of methods of direct communication of LIBRAS through current digital technologies. Córdova [2018] points out that while in Braille there is a direct transcription of written text into tactile text, for example, assistive translation technologies for hearing impaired people (such as subtitles and text boxes) are transcribed from spoken language to written language, and not directly to LIBRAS.

Corrêa *et al*. [2014] and Costa *et al*. [2024] discuss the existence of different applications for translating spoken Portuguese into LIBRAS, named HandTalk, ProDeaf and Pará-LIBRAS, and highlight their mediating and inclusive nature. However, the technological gap lies precisely in the opposite translation, when interpreting a LIBRAS sign and translating it into written or spoken Portuguese. Córdova [2018] and Quadros and Karnopp [2007] point out that textual adaptations cannot fully convey the meaning of a LIBRAS sign, as it consists of five phonological parameters: 1) facial expression; 2) hand orientation; 3) hand movement; 4) gesture localization; 5) hand configuration. Among these, the hand movement parameter cannot be expressed using static images. Furthermore, this language has a very strong inherited regional aspect, having "slangs" and regionalisms that vary according to the signer's experiences.

Given the needs identified by these studies, human-machine interaction technologies based on video detection are being explored. Figueiredo *et al*. [2025] applies image processing with MediaPipe on a proprietary mixed dataset of videos and images for the detection of dynamic LIBRAS signs using a hybrid CNN-LSTM architecture, considering not only a single static image of the sign, but the entire sequence of video frames that describe it. In their work, Figueiredo *et al*. [2025] achieved a performance of approximately 84% in classifying 6 LIBRAS signs. The present article uses a similar pipeline of dataset construction, image processing with MediaPipe, and real-time classification, but with a dataset containing more videos and a larger vocabulary, and an architecture based solely on CNN (convolutional neural networks) which, despite being simpler, showed better performance in the tests.

Badhe and Kulkarni [2015] also applied signal detection on a dataset created by the authors to translate Indian Sign Language signs into English using hand tracking and template matching, achieving a system accuracy close to 97.5%. Another author who works with its own dataset is Carvalho *et al*. [2021], who propose HandArch, a novel architecture for real time hand pose recognition from video to accelerate the development of sign language recognition application.

Tools like this are particularly interesting because they take into account the hand movement parameter, which cannot be processed using only image analysis. However, Both authors worked with datasets created entirely or partially by themselves. This occurs because of the lack of large and standardized video datasets for sign language. In general, even datasets with large vocabularies have few videos per category, which makes the real-time application of the network difficult. This article followed the same path, but with its own methodology for processing the data recorded in the dataset, in addition to final tests with real-time video captures with subjects never before seen by the network to prove the generalization and impersonality of the network's prediction.

As stated by Sarmento and Ponti [2023], researchers typically build their own video datasets in order to study the Sign Language classification problem. The limitations of such approach include having models that deal with a small vocabulary (number of words/categories) as well as a dependency on having a controlled environment, as also pointed by Rastgoo *et al*. [2021].

In this case, it is convenient to study the different aspects of building personal datasets for sign language detection and how to improve the generalization and democratization of those datasets. So, the present article follows some of the most prominent datasets in the literature. Li *et al*. [2020] propose the creation of Word-Level American Sign Language (WLASL), a new large-scale video dataset of American Sign Language (ASL), in addition to comparing the training efficiency of two different models when applied to this dataset: 1) holistic visual appearance based approach; 2) 2D human pose based approach. This dataset presents one of the largest vocabularies, featuring 2000 common different words in ASL performed by over 100 signers, however this dataset have an average 10.5 videos per category.

Almeida *et al*. [2019] published MINDS-LIBRAS, a dataset with RGB-D videos, containing 1) RGB videos; 2) videos with depth information; 3) information from 25 points/joints of the body; 4) 1347 points on the signer's face to analyze different neural network architectures to recognize LIBRAS signs. This dataset is recorded by 12 signers, each one performing 5 videos for each of the 20 words proposed, resulting in a 1200 video dataset. However, not all signers performed all signals. The lack of some videos makes the training unbalanced in certain signs.

Another observed dataset is the V-LIBRASIL, built by Rodrigues and Zanchettin [2024]. This LIBRAS dataset is composed of 1364 terms extracted and translated from the American Sign Language Lexicon Video Dataset (ASLLVD). Despite the large vocabulary, the V-LIBRAS dataset only has 3 videos per sign (each one captured by one different signer), which is very little for training a network to continuously classify signs in real-time.

Observing these cases, the present project sought its own architecture of dataset with fewer categories (20 signs), but

with more videos per category compared to the WLASL and V-LIBRASIL datasets and with a standardized number of videos per category.

To process the created dataset, this project studied the use of CNNs to analyze data sequences in the context of gesture and signal classification. Authors like Peixoto *et al*. [2021] and Peixoto *et al*. [2022] developed several algorithms to compare joint coordinates feature processing methods for gesture recognition with CNN and RNN in the MSRC-12 Dataset, applying a Kinect sensor camera to detect full-body gestures. The present article used simpler techniques applied to a similar problem, processing joint coordinates features with a CNN-based algorithm.

Rocha *et al*. [2020] propose a CNN-based algorithm to translate LIBRAS signs into Brazilian Portuguese, obtaining a precision of 90%. As a step forward in the application of this technique, the present project sought to remove the influence of the individual performing the signs. by aplying the MediaPipe algorithm while precessing the data. The present study acknowledges that, as attested by these authors and by Goodfellow *et al*. [2016], CNNs have great potential to perform tasks that require mapping an input vector to an output vector (such as classifying videos, images, or data sequences).

In addition, one of the main proposals of the present study is to develop an algorithm that could be used in several different environments. Similarly to what was done by De Oliveira Feliciano *et al*. [2023], which presents a viable methodology for the recognition of static and dynamic expressions of the LIBRAS, in environments whose background is complex.

3 Theoretical Background.

As mentioned in the previous section, Córdova [2018] and Carvalho *et al*. [2021] state that accessibility technologies for LIBRAS must take into account all five parameters of this language, including movement. Thus, video processing tools that extract only the fundamental features for recognizing the signs become particularly interesting.

3.1 MediaPipe Algorithm.

Mediapipe is an open source library developed by Lugaresi *et al*. [2019] for building solutions related to media perception and detection based on Neural Networks and Machine Learning. It consists of three main parts: a framework for sensory data interfacing, a set of tools for performance validation, and a collection of reusable interface and processing components, called calculators.

Among all the solutions provided by MediaPipe, this work uses only MediaPipe - Hands and MediaPipe - Pose, which are responsible for detecting and mapping hands and the pose of individuals in images, respectively. MediaPipe - Hands is a high-precision solution for detecting hands and fingers, employing machine learning to infer a map of 21 points in a three-dimensional space from a single frame. Similarly, MediaPipe - Pose creates a map of 33 points for the body pose of the individual in the camera. Figure 1 represents the configuration of the point map generated by MediaPipe - Hands.

According to its documentation, MediaPipe - Hands employs a pipeline composed of multiple models working together. A palm detection model operates on the entire image and returns a bounding box for the orientation of the detected palms. A hand identification model then operates on the image cropped by the first operation, returning a high-precision three-dimensional point map. MediaPipe - Pose works similarly, but it extracts points from the individual's posture in the frames (torso, shoulders, neck, etc.). Figure 2 exemplifies the creation of point maps by MediaPipe Hands in frames of various hand signs.

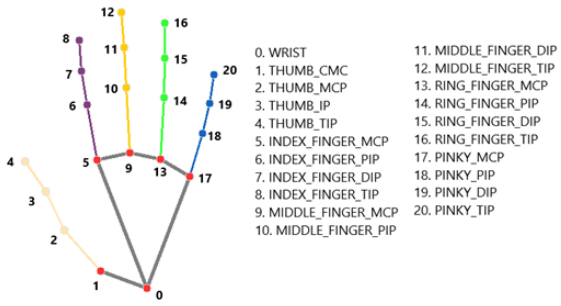

Figure 1. Arrangement of points generated by MediaPipe - Hands

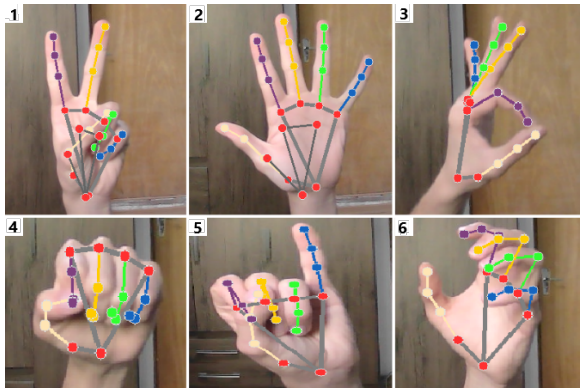

Figure 2. Examples of appliance of MediaPipe - Hands

The creation of these point maps frees the network from any interference from the scene or the physical characteristics of the people in the videos, since the resulting point matrix from this processing contains only the coordinates of the keypoints which represents the position of the individual's joints in a frame, and their movement throughout the frames. Also, when two people perform LIBRAS signs at different speeds, the number of frames in the videos is also different, so, in conjunction with Mediapipe, it was necessary to study ways to temporally standardize the videos used, since they do not have the same length.

3.2 Fast Dynamic Time Warping.

Fast Dynamic Time Warping (FastDTW) is a technique proposed by Salvador and Chan [2007] to find the best alignment between two series with different time durations. Figure 3 demonstrates this technique by illustrating two sequences with different dimensions on the x-axis, represented by the horizontal lines. Because FastDTW only standardizes the size of the sequences (x-axis), and not the information contained in each point of this sequence (y-axis), the values on the y-axis are irrelevant for this example. Note that since the upper sequence has twice as many points as the lower sequence, the perfect alignment would be one where each point "p" corresponds to

two points "P".

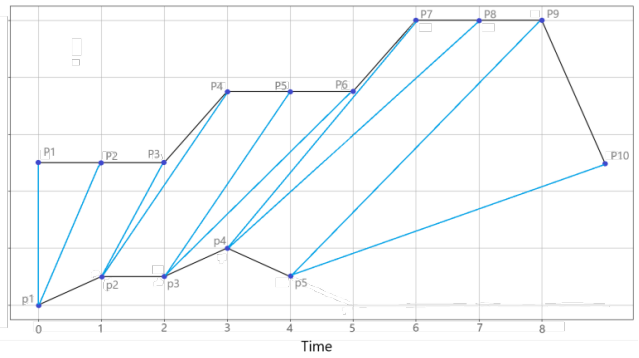

Figure 3. Representation of Dynamic Time Warping for two sequences

The Dynamic Time Warping problem can be described as: Given two temporal series X and Y of length $|X|$ and $|Y|$,

$$\begin{aligned} X &= x_1, x_2, ...x_i, ...x_{|X|} \\ Y &= y_1, y_2, ...y_i, ...y_{|Y|} \end{aligned} \tag{1}$$

Build a path of alignment W,

$$\begin{aligned} W &= w_1, w_2, ..., w_K \\ max(|X|, |Y|) &\leq K < |X| + |Y| \end{aligned} \tag{2}$$

Where K is the length of the path and the K-th element of this path is

$$w_k = (i, j) \tag{3}$$

Such that "i" is the index of the temporal series X, and "j" is the index of the temporal series Y. The optimal warp path is a warp path W with minimum distance (typically Euclidean distance). The distance of W is:

$$Dist(W) = \sum_{k=1}^{k=K} Dist(w_{ki}, w_{kj}) \tag{4}$$

Where $Dist(w_{ki}, w_{kj})$ is the distance between the two data point indexes (one from X and one from Y) in the k-th element of the warp path.

To solve this problem, Salvador and Chan [2007] propose a new way of approaching the issue. Instead of trying to solve the entire problem at once, they address smaller partitions of this larger problem. In this case, small partitions of the temporal sequences are observed, and the solutions found are applied to the whole problem. Then, they construct a two-dimensional cost matrix D, of size $|X|$ by $|Y|$, such that the value of $D(i, j)$ is the alignment path with the smallest possible distance that can be built from the two temporal series $X' = x_1...x_i$ and $Y' = y_1, ..., y_j$. The value of $D(|X|, |Y|)$ will contain the path with the smallest possible distance between the series X and Y.

If the path passes through all the cells $D(i, j)$ in the cost matrix, it means that the i-th point of series X is aligned with the j-th point of series Y. Since a point from one series can align with multiple points from the other series, they do not need to be of the same length.

This alignment created for the two temporal series can be used to find corresponding regions between them or determine their similarity. In this project, FastDTW is used to expand a smaller temporal series until it matches the size of a larger series, while maintaining the alignment between the smaller series before and after its expansion. The goal is for all videos in the LIBRAS dataset to have the same duration and, consequently, the same number of frames, by temporally expanding the shorter videos through the insertion of repeated frames, without losing their essence.

4 Methodology.

For the development of this research, the following steps were taken: 1) Dataset construction; 2) Application of the Mediapipe algorithm; 3) FastDTW data standardization; 4) Neural network training; 5) Real-time sign detection.

4.1 Dataset propose.

The decision to develop a custom dataset comes from the lack of large and standardized datasets in the literature for the Brazilian Sign Language. The Word-level American Sign Language dataset (WLASL), developed by Li *et al.* [2020], for example, has a vocabulary of over 2,000 American Sign Language signs, but has an average of only 10 videos per sign. Some LIBRAS datasets like V-LIBRASIL from Rodrigues and Zanchettin [2024] are even smaller. The MINDS-LIBRAS dataset, developed by Almeida *et al.* [2019], on the other hand, have 60 videos per sign (5 videos recorded by each one of the 12 signers). However, it is not standardized, which means that some videos are missing because a few signers did not recorded some videos.

Following the structure proposed by Almeida *et al.* [2019], this study focuses on a vocabulary of 20 LIBRAS signs, which are the equivalents of the words "Hug", "Help", "Joy", "Friend", "Good", "to Play", "House", "Start", "to Eat", "Day", "Happy", "LIBRAS", "Thank you", "Hello", "Stop", "Please", "Teacher", "Bad", "to Know" and "Deaf".

The selection of the words that compose the dataset took into account gramatical variation in Portuguese and the structure of the corresponding signs in LIBRAS. From the Portuguese perspective, the dataset includes verbs, adjectives, nouns, and interjections commonly used in everyday Brazilian contexts. From the LIBRAS perspective, the selected signs present a wide variation across the five parameters that constitute signs, encompassing different facial expressions, hand orientations, hand movements, gesture localizations, and hand configurations.

For the creation of this dataset, 400 videos were captured, each 6 seconds long, with 20 videos for each of the proposed signs. All videos are recorded with an built-in 1080p camera from a Lenovo i7 notebook, with the same person performing the signs. Figure 4 shows some frames taken from the videos where the signer performs the signs for 'Friend' (A), 'Please' (B), and 'Thank you' (C).

The decision to build a dataset composed of videos from only one signer was made due to the lack of time and resources to assemble a more complex dataset. It is known that this choice may influence the final performance of the neural network, making the classification biased by the signer's appearance, background, and the way the signs are performed.

Figure 4. Sign Frames taken from the author's dataset

Throughout the project, efforts were made to apply algorithms that process the videos in order to mitigate these biases. The following sections explain in greater detail how this was accomplished and how the dataset format influenced the final prediction results.

With the videos captured, an initial segmentation of the videos is performed using the numpy and opencv libraries, separating them into sequences of frames. Then, to reduce the amount of unnecessary data at the beginning and end of the videos, frames where the signer has their hands down are removed, leaving only the sign itself to be processed. Note that the final dataset is not composed by the captured videos, but by the numpy matrix resulting from the processing of those videos by MediaPipe and FastDTW algorithms. This processing is detailed in the subsequent sections.

4.2 Application of Mediapipe algorithm.

Once the frames are separated, the MediaPipe algorithm can be applied to all frames of each video to extract the relative position of the signer's joints in each one. This algorithm extracts the x, y and z coordinates of each of the 86 joint points in the image. Of these points, 21 represent the joints of the left hand, 21 of the right hand, and the remaining 44 are from the signer's pose (arms, legs, torso, shoulders, etc.). Figure 5 shows the general dimensions of the dataset, composed by 20 videos for each sign, where each row in a video represents a frame from the video and each column represents an coordinate (x, y, and z) of each point generated by Mediapipe, totaling 258 columns.

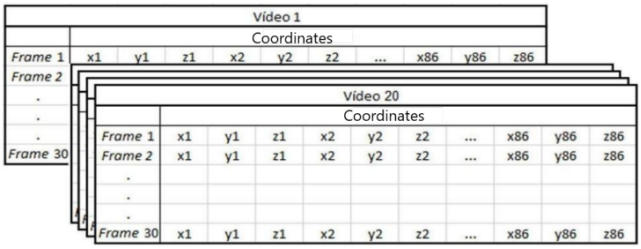

Figure 5. Architecture of data for each sign

Figure 6 also shows a representation of how the Mediapipe-generated point map is placed in the three-dimensional space according to the parts of the signer's body while performing the sign for "Day".

Thus, the final stored dataset is composed solely by the sequences of point maps generated by Mediapipe, which represent each performed sign. Figure 7 exemplifies the final result of the video processing from the perspective of the neural network. It is noted that the network only perceives the point maps that describe the movements made by the signers in the videos. Figure 7 is only a visual representation of the keypoints, the resulting data is a numpy matrix with the coordinates of each point, not an image.

Figure 6. Mediapipe Pose and Hands printed in comparison over a frame

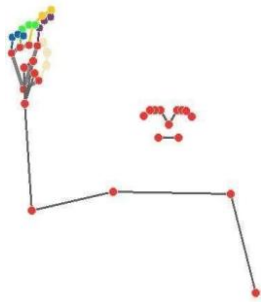

Figure 7. Point arrangement generated by MediaPipe

4.3 FastDTW data standardization.

The next step in data standardization concerns the video dimensions after the previously described cuts. In Figure 5, it can be seen that each video has 30 frames. This number of frames is a result of the FastDTW standardization. As mentioned earlier, each signer performs a LIBRAS sign movement at a different speed, resulting in videos with varying durations and, consequently, a different number of frames. It is essential for the network to function properly that all input data have the same dimensions. For this standardization, the Fast Dynamic Time Warping (FastDTW) algorithm was used, which concatenates sequences with different temporal dimensions into the same format, following an "optimal" alignment for both sequences.

After this processing step, the input data for the network is correctly standardized for the neural network's operation. There are 20 different signs, each with 20 videos of 30 frames each. These frames are numerically represented by vectors with the x, y, and z coordinates of each of the 86 points on the map generated by Mediapipe, totaling 258 values.

4.4 Neural Network training.

Finally, the dataset is divided into two groups. One contains the training data, which is passed to the neural network to train it to recognize and differentiate the signs, and the other is composed of validation data, which is used to verify the effectiveness of the already trained neural network. The data was split into 70% for training and 30% for validation for each sign.

This work aimed to design a convolutional neural network with 4 internal layers, one input layer, and one output layer, with two internal layers having 3x3 kernel convolutions and two fully connected layers (also called "dense layers"). All these layers use the "ReLU" activation function. At the end of the training process, the network is expected to distinguish different signs based on the similarity to the data

presented to it during training.

The data from the processing stages is then fed into this network during the training phase. In each iteration, a video passes through the network, updating its weights and biases to reduce the error between the output predicted by the network and the known output until the entire dataset has been processed by the network. This learning process is repeated for a predefined number of training epochs or until the accuracy exceeds a desired value. Fifty training epochs were applied, using the "Adam" optimizer, "categorical cross-entropy" loss, and evaluation metrics based on "categorical accuracy".

4.5 Real-time sign detection

Finally, Python algorithms for real-time image capture are applied to build frame sequences similar to those used in the training and validation of the network. These sequences are assembled in real-time and updated frame by frame as the computer's camera adds new frames. Using the last n images, where n is the number of frames used to train the network, the already trained model is used to attempt to classify the executed LIBRAS sign, aiming to validate the effectiveness of the proposed methodology.

5 Results.

Initially, the results obtained from processing the video frames with the Mediapipe tool are presented. The use of this algorithm is particularly interesting because it removes interference from all the noise generated by the video recording environment, such as lighting, background, or the presence of other people in the scene, as well as any interference from the physical characteristics of the person performing the sign, such as gender, skin tone, and body type. Figure 8 demonstrate the correct functioning of Mediapipe for people with different characteristics and in different environments.

Figure 8. MediaPipe processing for videos of different people in different enviroments

Occasional failures in the creation of the point maps occurred only when the videos were recorded against the sun, which greatly increases the ambient brightness, or when the signer performed the sign too quickly or too close to the camera. These scenarios are rare and, considering the scope of the proposed tool, inevitable in this context.

Regarding the network training, the accuracy quickly converged to 99.9%, reaching its peak by the 23rd training epoch as can be seen in Figure 9.

The 30% of the dataset that was not used to train the network was passed to the neural network to evaluate its efficiency. From these tests with this segment of the original dataset, it was possible to generate the confusion matrix that demonstrates the characteristic of the error found for each specific signal. In addition to this, accuracy, recall and F1-score metrics were determined for each signal. These metrics are

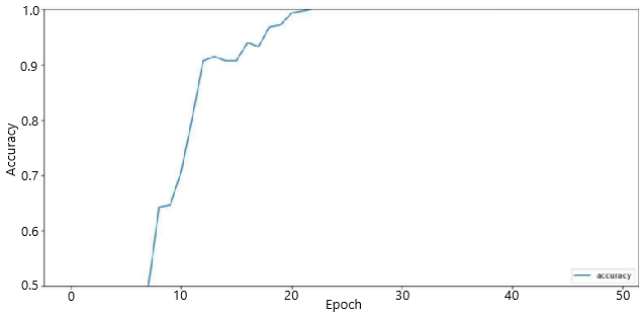

Figure 9. Network accuracy for each training epoch

cataloged in the table 1, while Figure 10 shows the confusion matrix generated.

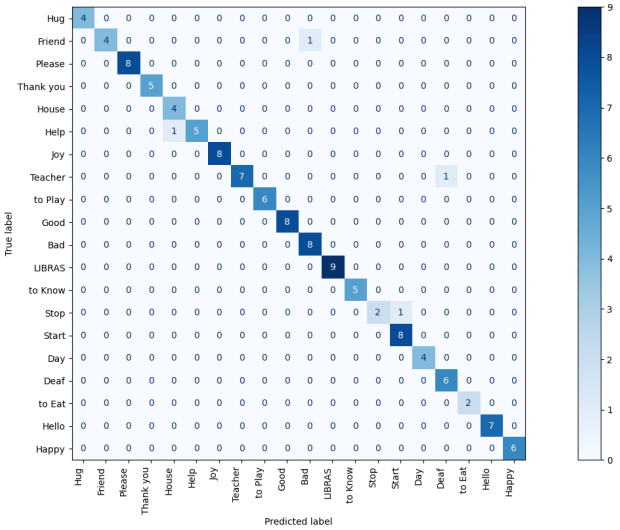

Figure 10. Confusion Matrix

Analyzing both Figure 10 and Table 1, it is observed that the model, when subjected to a separate test plot from the initial project dataset, obtained excellent prediction results. The overall accuracy of the test was approximately 91.6%, demonstrating the network's proficiency in categorizing the presented signals. However, the learning curve shown in Figure 9 and the results close to 100% accuracy, recall, and precision may indicate overfitting of the network, which occurs when the trained model only "memorizes" the data instead of actually learning from them. In this work, given the proposed dataset architecture, we seek to avoid overfitting as much as possible, always trying to achieve a generalized and standardized network that can detect signals made by any future user, and not just those signals present in the training dataset.

Considering this context and the need to prove the network's functionality beyond the initial database, we subjected the model to real-time testing, separating a test group with individuals different from those used to build the dataset. Each member of the test group performed LIBRAS signs in front of the same camera used to capture the dataset. The algorithm recognizes when the signer has their hands in front of the camera by processing the incoming frames with MediaPipe, and if a sequence of more than 5 frames is detected, it recognizes that a sign is being made. When the hands go down, the sequence of frames generated is passed to FastDTW

Table 1. Precision, Recall e F1-score per class

Class	Precision	Recall	F1-score	Support
Hug	1.0000	1.0000	1.0000	8
Friend	1.0000	0.7143	0.8333	7
Please	1.0000	1.0000	1.0000	4
Thank you	1.0000	1.0000	1.0000	7
House	0.8000	1.0000	0.8889	4
Help	1.0000	0.8333	0.9091	6
Joy	1.0000	1.0000	1.0000	4
Teacher	0.8333	1.0000	0.9091	5
To Play	0.7500	1.0000	0.8571	6
Good	0.4286	0.7500	0.5455	4
Bad	0.8750	0.7778	0.8235	9
LIBRAS	1.0000	1.0000	1.0000	9
To Know	1.0000	1.0000	1.0000	4
Stop	1.0000	0.7778	0.8750	9
Start	1.0000	1.0000	1.0000	7
Day	1.0000	1.0000	1.0000	7
Deaf	0.8333	0.8333	0.8333	6
To Eat	1.0000	1.0000	1.0000	5
Hello	1.0000	0.7500	0.8571	4
Happy	1.0000	1.0000	1.0000	5

to standardize the length, and finally imputed to the trained model to classify the sign. That way, the frames captured in real-time undergo the same processing necessary to fit the neural network's input (MediaPipe and FastDTW processing) and are then shown to the trained model, which attempts to predict which sign is being performed in the frames presented to it.

Thus, each sign was tested 5 times by each of the 8 signers in the test group, totaling 40 repetitions per sign. Regarding the diversity of the group, it was composed of 6 men and 2 women aged between 14 and 43 years. Only one individual is Black, and the rest are White. One individual was born and resided most of their life in the southeastern region of Brazil, the others are from the southern region. With the exception of the author of the project, none of the individuals had prior knowledge of LIBRAS, and it is difficult to measure the mannerisms and "accent" of each of them in the performance of the signs, but it is evident that, given the highly personal nature of communication, each individual performed the signs with their specific particularities. The neural network's classification accuracy rate for each of the 20 proposed signs is described in Figure 11.

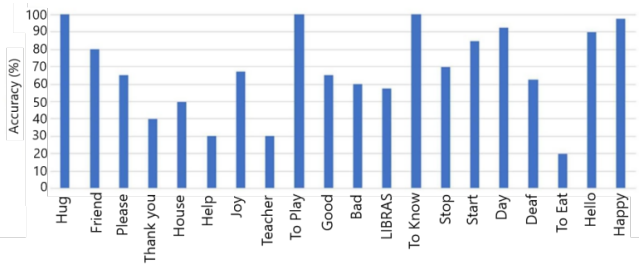

Figure 11. Overall network accuracy for each sign

It is noted that the network's accuracy varies significantly depending on the sign being tested. Some signs, such as "Hug", "Play", and "Know", were correctly understood by the network in all tests, while "Help", "Eat" and "Teacher" showed less than 40% accuracy.

It was observed that, regardless of the person performing the sign, the background setting, or the distance from the camera, the results remained the same, indicating that these parameters did not influence the network's correct understanding of the signs. This is attributed to the effectiveness of the data processing methodology presented earlier, primarily due to the processing of videos by MediaPipe. The conversion of videos into NumPy matrix sequences containing only information about the movement and position of the hand, arm, and torso joints of the signers were made with similar success of studies like those from Lugaresi *et al*. [2019], Bhamidipati *et al*. [2023] and Figueiredo *et al*. [2025], and exempted the final network classification from parameters linked to the signaler's individual physical characteristics (such as skin color, gender, and age). Thus, we can conclude that the data treatment and standardization processes successfully eliminated the influence of unwanted factors, such as the physical characteristics of the test subjects and the video setting.

The factors that hindered detection were related to how the test subjects performed the signs, with variations in location and movement parameters. The negative results for some signs are due to two main factors: 1) their complexity or similarity to other signs; 2) the difference between how each signer performed the sign when compared to the dataset performances. In both cases, the bad real-time results for some signs demonstrate the need to use a larger and more varied database. A dataset built by only one signer limits the classification to the mannerisms and "accent" of the dataset. It is also evident that the network struggles to distinguish subtle nuances in similar signs when trained with only 20 videos per sign.

This result is similar for the one obtained by Figueiredo *et al*. [2025], who, despite achieving about 84% of accuracy in sign classification, claims that the most frequent confusions occurred between signals with similar manual configurations, reinforcing the need to expand the database to better discriminate between these classes.

Even so, the fact that the network is able to identify most signals with an effectiveness equal to or greater than 60% Thus, the overall result was satisfactory for most of the proposed signs, achieving the main objectives of this study.

Future works will focus on enhancing the dataset by collecting data from various signers, in more quantity and with a larger vocabulary, building the data with a more defined method. Furthermore, the application of different, more complex neural networks will be studied, mainly with the development of larger CNNs and their comparison with long-term memory networks (LSTM).

6 Conclusion.

This study highlights the importance of developing new accessibility tools for hearing-impaired and for LIBRAS speakers, while successfully proposing a methodology for the detection and classification of continuous signs in Brazilian Sign Language using MediaPipe, FastDTW and convolutional neural networks. This improves the integration of communication, information, entertainment, and education technologies and electronic devices with the hearing-impaired community.

Clearly, the final result requires refinement, and future work will focus mainly on increasing the dataset with videos of various people, as the low number of data was a key factor in the low accuracy for detecting more complex signs. Nonetheless, the results were promising and demonstrated the functionality and versatility of the methodology on detecting 20 different LIBRAS signs in real-time using the proposed architecture thus achieving the initial objectives of the project.

Declarations

Acknowledgements

We acknowledge the Universidade Federal de Santa Maria students and teachers who participated in this research. ChatGPT, an artificial intelligence tool, were used to support the review of this article's English grammar.

Funding

This work was carried out with the technical and financial support of the Coordination for the Improvement of Higher Education Personnel (CAPES).

Authors' Contributions

Luciano Cardona worked on software development, the application of the proposed methodology, data collection and analysis, as well as the writing and translation of the article, under the guidance and supervision of Daniel Gamarra. Anselmo Cukla, Solon Bevilacqua and Daniel Bernardon contributed with insights into the results, reviewing, editing, correcting, and offering suggestions to improve this document. All authors have read and approved the final manuscript

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The materials used in this research are available in a github repository, at the link https://github.com/cardonixjr/CNN-for-LIBRAS. This availability enhances consultation, reproducibility, and, consequently, the principles of open science and knowledge sharing.

Ethical Issues

The authors acknowledge that the study involves data collection from a group of individuals. Therefore, it is important to clarify that all data collected in this study are preprocessed before being stored. The dataset consists only of numerical matrices of data already processed by MediaPipe and FastDTW, which do not contain detailed information about the appearance or identity of the captured individual. The same occurs in real-time tests, which are based only on the comparison between numerical matrices that only contain information about the relative position of the individual in the image and the movement of their body, making individual identification impossible from this raw data.

Regarding the real-time tests, it's important to clarify that the videos recorded during this stage are never stored. The tests are performed using continuous video streams, without saving any images, and the corresponding results are computed during the execution of the algorithm itself, without the need for post-processing or information recovery.

Also, in figures 4, 6 and 8, the individual appearing in the images is the own author of the work, who is the same individual that recorded the main dataset of this project and consents to the use of their image and data for publication.

Finally, all study participants signed an informed consent form guaranteeing their right to decide on the use of their data, transparency regarding the development and publication processes of the study, and full privacy and security against any data leakage throughout the duration of the research. The document also states that the produced dataset will be deleted after five years, ensuring that the data will not be permanently retained. Thus, the authors declare that all appropriate measures were taken to ensure the security, anonymity, privacy, transparency, and decision-making autonomy of the study participants.

References

Abdallah, E. E. and Fayyoumi, E. (2016). Assistive technology for deaf people based on android platform. *Procedia Computer Science*, 94:295–301. DOI: https://doi.org/10.1016/j.procs.2016.08.044.

Almeida, S. G. M., Rezende, T. M., Almeida, G. T. B., Toffolo, A. C. R., and Guimarães, F. G. (2019). Minds-libras dataset. Zenodo. DOI: https://doi.org/10.5281/zenodo.2667329.

Antunes, D. R., Guimarães, C., García, L. S., Oliveira, L. E. S., and Fernandes, S. (2011). A framework to support development of sign language human-computer interaction: Building tools for effective information access and inclusion of the deaf. In *2011 Fifth International Conference on Research Challenges in Information Science*, pages 1–12. DOI: https://doi.org/10.1016/10.1109/RCIS.2011.6006832.

Badhe, P. C. and Kulkarni, V. (2015). Indian sign language translator using gesture recognition algorithm. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, pages 195–200. DOI: https://doi.org/10.1109/CGVIS.2015.7449921.

Bastos, I. L., Angelo, M. F., and Loula, A. C. (2015). Recognition of static gestures applied to brazilian sign language (libras). In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 305–312. DOI: https://doi.org/10.1109/SIBGRAPI.2015.26.

Bhamidipati, V. S. P., Saxena, I., Saisanthiya, D., and Retnadhas, M. (2023). Robust intelligent posture estimation for an ai gym trainer using mediapipe and opencv. In *2023 International Conference on Networking and Communications (ICNWC)*, pages 1–7. DOI: https://doi.org/10.1109/ICNWC57852.2023.10127264.

Brasil (2005). Decreto nº 5.626, de 22 de dezembro de 2005. Diário Oficial da União. Regulamenta a Lei nº 10.436, de 24 de abril de 2002, que dispõe sobre a Língua Brasileira de Sinais - Libras, e o art. 18 da Lei nº 10.098, de 19 de dezembro de 2000.

Brasil (2015). Lei nº 13.146, de 6 de julho de 2015. Diário Oficial da União. Institui a Lei Brasileira da Inclusão da Pessoa com Deficiência. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm. Accessed on 07 July 2026.

Carvalho, G., Brandão, A., and Ferreira, F. (2021). Handarch: A deep learning architecture for libras hand configuration recognition. In *Anais do XVII Workshop de Visão Computacional*, pages 19–24, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/wvc.2021.18883.

Castro, S. S., Lefèvre, F., Lefèvre, A. M. C., and Cesar, C. L. G. (2011). Acessibilidade aos serviços de saúde por pessoas com deficiência. *Revista de Saúde Pública*, 45(1):99–105. DOI: https://doi.org/10.1590/S0034-

89102010005000048.

Corrêa, Y., Maristela, C., Santarosa, L., and Biazus, M. C. (2014). Aplicativos de tradução para libras e a busca pela validade social da tecnologia assistiva. In *XXV Simpósio Brasileiro de Informática na Educação*, page 164. DOI: https://doi.org/10.5753/cbie.sbie.2014.164.

Costa, A. P., Diniz, E., Sales, E., Coelho, F., Costa, W. C., and Sousa, L. (2024). Usabilidade de um aplicativo de libras para educação e governança em uma cidade amazônica. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 1248–1262, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/sbie.2024.242459.

Courtin, C. (2000). The impact of sign language on the cognitive development of deaf children: The case of theories of mind. *The Journal of Deaf Studies and Deaf Education*, 5(3):266–276. DOI: https://doi.org/10.1093/deafed/5.3.266.

Córdova, G. D. (2018). Língua de herança: Língua brasileira de sinais – livro de ronice müller de quadros. *Pensares em Revista*, 13. DOI: https://doi.org/10.12957/pr.2018.33402.

De Oliveira Feliciano, F. D., De Andrade Briano, A., Prudencio, R. B. C., and Alves, E. C. (2023). Recognition of static or dynamic libras words in complex background environments. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–4. DOI: https://doi.org/10.23919/CISTI58278.2023.10211498.

Dias, T. S., Junior, J. J. A. M., and Pichorim, S. F. (2024). Classification of brazilian sign language gestures based on recurrent neural networks models, with instrumented glove. In Marques, J. L. B., Rodrigues, C. R., Suzuki, D. O. H., Marino Neto, J., and García Ojeda, R., editors, *IX Latin American Congress on Biomedical Engineering and XXVIII Brazilian Congress on Biomedical Engineering*, pages 611–620, Cham. Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-49407-9_61.

Figueiredo, A., Guarisi, A., Valinho, C., Marcelino, G., and Souza, V. (2025). De sinais a texto: Um sistema inteligente para tradução automática de libras com foco na inclusão educacional. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, pages 1489–1496, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/sbie.2025.12170.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA. ISBN: 978-0262035613.

Hommes, R. E., Borash, A. I., Hartwig, K., and DeGracia, D. (2018). American sign language interpreters perceptions of barriers to healthcare communication in deaf and hard of hearing patients. *Journal of Community Health*, 43:956–961. DOI: https://doi.org/10.1007/s10900-018-0511-3.

Instituto Brasileiro de Geografia e Estatística (2019). Pesquisa nacional de saúde. https://www.ibge.gov.br/estatisticas/sociais/saude. Accessed on 07 July 2026.

Li, D., Opazo, C. R., Yu, X., and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1448–1458. DOI: https://doi.org/10.1109/WACV45572.2020.9093512.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M. G., Lee, J., Chang, W., Hua, W., Georg, M., and Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172. DOI: https://doi.org/10.48550/arXiv.1906.08172.

Palavissini, C. F. C., Lima, K. R. L. d., Castro, L. P. V. d., and Lima, D. F. d. (2021). Digital information and communication technologies on the acquisition of scientific knowledge to deaf students: an integrative literature review. *Research, Society and Development*, 10(16). DOI: https://doi.org/10.33448/rsd-v10i16.23998.

Peixoto, J. S., Cukla, A. R., Welfer, D., and Gamarra, D. F. T. (2022). Comparison of different processing methods of joint coordinates features for gesture recognition with a rnn in the msrc-12. In Abraham, A., Gandhi, N., Hanne, T., Hong, T.-P., Nogueira Rios, T., and Ding, W., editors, *Intelligent Systems Design and Applications*, pages 498–507, Cham. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-96308-8_46.

Peixoto, J. S., Pfitscher, M., de Souza Leite Cuadros, M. A., Welfer, D., and Gamarra, D. F. T. (2021). Comparison of different processing methods of joint coordinates features for gesture recognition with a cnn in the msrc-12 database. In Abraham, A., Piuri, V., Gandhi, N., Siarry, P., Kaklauskas, A., and Madureira, A., editors, *Intelligent Systems Design and Applications*, pages 590–599, Cham. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-71187-0_54.

Quadros, R. M. D. and Karnopp, L. B. (2007). *Língua de sinais brasileira: estudos lingüísticos*. Biblioteca Artmed. Lingüística. Artmed, São Paulo. ISBN: 9788536303086. http://bds.unb.br/handle/123456789/948. Accessed on 11 July 2026.

Rastgoo, R., Kiani, K., and Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794. DOI: https://doi.org/10.1016/j.eswa.2020.113794.

Rocha, J., Lensk, J., Ferreira, T., and Ferreira, M. (2020). Towards a tool to translate brazilian sign language (libras) to brazilian portuguese and improve communication with deaf. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–4. DOI: https://doi.org/10.1109/VL/HCC50065.2020.9127257.

Rodrigues, A. and Zanchettin, C. (2024). V-librasil - a new dataset with signs in brazilian sign language (libras). IEEE Dataport. DOI: https://doi.org/10.21227/v589-hn68.

Rumjanek, V. and da Silva, W. (2019). Ciência para todos? *Revista Brasileira de Pós-Graduação*, 15:1–20. DOI: https://doi.org/10.21713/rbpg.v15i34.1606.

Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580. DOI: https://doi.org/10.3233/IDA-2007-11508.

Sarmento, A. H. d. A. and Ponti, M. A. (2023). A cross-dataset study on the brazilian sign language translation. In *IEEE/CVF International Conference on Computer Vision Workshops - ICCVW*. IEEE. DOI:

https://doi.org/10.1109/ICCVW60793.2023.00300.

Silva, D., Araujo, T., Rêgo, T., Brandão, M., and Gonçalves, L. (2023). A multiple stream architecture for the recognition of signs in brazilian sign language in the context of health. *Multimedia Tools and Applications*, 83. DOI: https://doi.org/10.1007/s11042-023-16332-7.

Stopa, S. R., Szwarcwald, C. L., Oliveira, M. M. d., Gouvea, E. d. C. D. P., Vieira, M. L. F. P., Freitas, M. P. S. d., Sardinha, L. M. V., and Macário, E. M. (2020). Pesquisa nacional de saúde 2019: histórico, métodos e perspectivas. *Epidemiologia e Serviços de Saúde*, 29(5):e2020315. DOI: https://doi.org/10.1590/S1679-49742020000500004.

Trotta, T., Rocha, L., de Andrade, T. R., de Paiva Guimarães, M., and Dias, D. R. C. (2022). C-libras: A gesture recognition app for the brazilian sign language. In Gervasi, O., Murgante, B., Hendrix, E. M. T., Taniar, D., and Apduhan, B. O., editors, *Computational Science and Its Applications – ICCSA 2022*, pages 603–618, Cham. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-031-10522-7_41.