applied sciences

MDPI

Article

Dual-Stream BiLSTM–Transformer Architecture for Real-Time Two-Handed Dynamic Sign Language Gesture Recognition

Enachi Andrei [1,2,*], Turcu Corneliu-Octavian [1], Culea George [2], Andrioaia Dragos-Alexandru [2], Ungureanu Andrei-Gabriel [1,2] and Sghera Bogdan-Constantin [1,2]

[1] Faculty of Electrical Engineering and Computer Science, "Stefan cel Mare" University of Suceava, University Street No. 13, 720229 Suceava, Romania; cturcu@usm.ro (T.C.-O.); andrei-gabriel.ungureanu-fom@ing.ub.ro (U.A.-G.); bogdan-constantin.sghera-frm@ing.ub.ro (S.B.-C.)

[2] Department of Energetics and Computer Science, Faculty of Engineering, "Vasile Alecsandri" University of Bacău, Mărășești No. 157, 600115 Bacău, Romania; gculea@ub.ro (C.G.); dragos.andrioaia@ub.ro (A.D.-A.)

* Correspondence: andrei.enachi@ub.ro

Abstract

Two-handed dynamic gesture recognition represents a fundamental component of sign language interpretation involving the modeling of temporal dependencies and inter-hand coordination. In this task, a major challenge is modeling asymmetric motion patterns, as well as bidirectional and long-range temporal dependencies. Most existing frameworks rely on early fusion strategies that merge joints, keypoints or landmarks from both hands in early processing stages, primarily to reduce model complexity and enforce a unified representation. In this work, a novel dual-stream BiLSTM–Transformer model architecture is proposed for two-handed dynamic sign language recognition, where parallel encoders process the trajectories of each hand independently. To capture spatial and temporal dependencies for each hand, an attention-based cross-hand fusion mechanism is employed, with hand landmarks extracted by the MediaPipe Hands framework as a preprocessing step to enable real-time CPU-based inference. Experimental evaluation conducted on custom Romanian Sign Language dynamic gesture datasets indicates that the proposed dual-stream-based system outperforms single-handed baselines, achieving improvements in high recognition accuracy for asymmetric gestures and consistent performance gains for synchronized two-handed gestures. The proposed architecture represents an efficient and lightweight solution suitable for real-time sign language recognition and interpretation.

Keywords: Romanian sign language; convolutional neural networks; transformer; MediaPipe; dynamic gestures; landmarks; patterns; BiLSTM; CPU; cross-hand

Academic Editor: Thomas Lindner

Received: 30 January 2026
Revised: 12 March 2026
Accepted: 13 March 2026
Published: 18 March 2026

1. Introduction

The Romanian Sign Language Alphabet (RSL) relies heavily on visual cues derived from spatial and temporal configurations, as well as dynamic motion trajectories, which serve as the primordial communication channel. Although this topic has gained increasing interest, the latest approaches are limited in their ability to process two-handed complex gestures that enable a natural way of communication with signers. By modeling spatial–temporal dependencies, the proposed system aims to improve accessibility to the latest technologies and public services, while addressing limitations.

Existing RSL recognition systems primarily focus on static keypoint classifiers, often convolutional encoder-based, which perform well on isolated datasets of gestures, but fail to capture dynamic motion phases and temporal dependencies from consecutive frames. Also,

many dynamic gesture recognition systems merge both hands in early processing while overlooking asymmetric motion patterns [1]. This design directly affects the recognition robustness of the model in real-world scenarios, where gestures vary in execution speed, signer style and real-world environments. Recently, Transformer-based architectures have demonstrated strong capabilities in modeling long-range temporal dependencies, such as long-range sequences of data. However, in architectures that rely on early fusion of both hands, the full potential of the Transformer model remains unutilized. This limitation leads to suboptimal spatial and temporal modeling and also ignores some essential patterns for two-handed gestures. To address this gap, a Dual-Stream BiLSTM–Transformer model architecture is proposed for two-handed dynamic gesture recognition in the RSL alphabet.

The proposed model processes each hand trajectory independently through parallel streams, followed by an attention-based cross-hand encoder. The encoder weights both hands according to their linguistic corresponding contribution. The asymmetric motions are captured between inter-hand coordinates, which enables the proposed system to build a solid structure or a pattern for specific temporal gestures. MediaPipe Hands is employed as the main framework, used for hand landmarks extraction, enabling real-time inference in real-world scenarios. The new architecture works on sign-language modeling and is evaluated through controlled experiments (variable environments) and an ablation study to investigate its effectiveness with dynamic gestures [2]. The main contributions of the proposed study are as follows:

- A new hybrid Dual-Stream Architecture that uses the fusions between BiLSTM and Transformer encoders to model both hands independently;
- A fusion cross-hand model that captures both dependencies, obtaining specific dynamic patterns;
- Generalization over signer inter-variability through landmark coordinate normalization in different working environments using MediaPipe;
- High accuracy with symmetric or asymmetric dynamic hand gestures [3].

In addition, the proposed system aims to provide a scalable and computationally efficient solution framework for real-time dynamic gesture recognition, representing an important step for RSL classification and interpretation. The proposed system will investigate and answer the following research questions:

- Will the proposed dynamic recognition system improve accuracy and overall performance compared to other early fusion approaches?
- Will the cross-hand Transformer-based approach improve classification and gesture recognition with asymmetric or asynchronous hand trajectories?
- Will the proposed system record better inference under unconstrained different working environments in real-world scenarios based on hand keypoints or landmarks?

Beyond the technical and overall performance, the proposed system is also designed to be an assistive companion, acting as a communication framework with a direct impact on society. In general, it aims to support and help members who utilize sign language communication, enabling interaction with new technologies. Additionally, by integrating gesture into text translation and vocal feedback, the proposed system can construct and elaborate in real-time grammatically correct sentences. This solution is not only robust for dynamic sequences of gestures, but also enables accessibility on human motion and computer interaction, removing communication barriers in real time. The proposed study focuses on a controlled subset of isolated two-handed dynamic gestures serving as a methodological validation of the proposed architecture rather than a comprehensive sign language vocabulary.

The following research is organized as follows: Section 2 presents recent approaches focused only on dynamic sign language gesture recognition emphasizing two-handed models and early fusion methods; Section 3 presents the entire architecture of the proposed recognition system, including the used datasets, data acquisition, hand keypoints and landmark preprocessing using MediaPipe Hands as the framework and the BiLSTM–Transformer model; Section 4 illustrates the experimental setup and the obtained quantitative results; Section 5 discusses the strong points, limitations and findings for the proposed recognition system; and the final section concludes the research and draws the future direction of the proposed study.

2. Related Work

Video-based recognition approaches are typically processed by bidimensional or tridimensional convolutional neural networks that capture spatial and temporal correlations from RGB sequences [4–7]. The tridimensional approaches use kernel correlations to capture gesture sequences, but only on very large and complex datasets. They often struggle to recognize trajectory and gesture variations and stay very sensitive to occlusions. However, optical flow approaches are effective mainly for complex gestures, but are strictly dependent on contrast and low-light illumination working environments. Regarding landmarks or articulations-based representations, optical flow lacks precision and fails to capture small movements across consecutive frames. Body and hand pose skeleton recognition approaches are also an alternative line of research that use small inputs to eliminate background noises. Systems often tend to use OpenPose or MediaPipe Hands frameworks to feed hand landmark coordinates into LSTMs or GCNs. Graph Convolutional Networks (GCNs) are widely used in skeleton-based gesture recognition because they model spatial dependencies between articulated joints and their temporal evolution [8–10]. This emergence will suppress characteristic motion patterns from the sign language alphabet (asymmetric and asynchronous gestures). This approach is limited in terms of temporal modeling, which cannot effectively interpret inter-hand coordination under temporally misaligned hand movements [11–15].

For dynamic interpretation, recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures have been widely used for modeling temporal dependencies in gesture recognition tasks [16,17]. To improve accuracy, sequential models employ two bidirectional variants, preserving future and past short-range context details. Rarely, RNNs use independent streams for both hands; instead, they emerge early into one trajectory, which increases model ambiguity for certain types of gestures where only one hand has a dominant semantic role [18,19]. The latest Transformer-based architectures recorded high performance in multiple gesture recognition tasks, but also in their ability to preprocess long-range spatial and temporal dependencies by using self-attention mechanisms [20,21]. For example, Camgoz et al. proposed a sign language Transformer framework for joint recognition and translation, demonstrating that encoders can model long-range dependencies directly from video sequences [22]. Similarly, De Coster et al. explored sign language recognition with a Transformer-based architecture, proving that self-attention mechanisms can capture complex temporal relationships between gesture frames. More recently, Kothadiya et al. introduced SIGNFORMER, a Transformer-based vision architecture design for sign language recognition using visual hand gesture representations. Although these approaches demonstrate strong performance in temporal modeling dependencies, many of them rely on RGB video sequences or fused gesture representations, which significantly increases computational complexity and may suppress asymmetric hand dynamics. In contrast, the proposed study focuses on landmark-based compact input gesture representations extracted with MediaPipe Hands, enabling real-time inference. Furthermore, the

proposed architecture models left- and right-hand trajectories using a dual-stream BiLSTM encoder, followed by a Transformer-based cross-hand fusion mechanism, which enables adaptive modeling of asymmetric and asynchronous two-handed gestures. They do not rely on sequential memory, which makes them well-suited for modeling extended gesture sequences and long-range temporal dependencies [23–26].

Early fusion refers to the concatenation of hand features at the input level prior to temporal modeling, enforcing implicit synchronization between hand trajectories [27]. They fuse both hands into a single representation trained on specific hand trajectories. This reduces the capability of the self-attention mechanism to process and build a general model from both hands that is able to interpret asymmetric two-handed dynamic gestures. A frequent limitation in sign language recognition systems is early fusion. This stage enables the system to concatenate both inputs from the right and left hand, reducing computational complexity [27]. In real-world scenarios, sign language recognition uses asynchronous gestures mirrored and independent of the hands.

Therefore, merging both hands into a single representation leads to a major bidirectional confusion between hands and sign semantics, reducing robustness under inter-signer variability. In skeleton-based approaches, dual-stream architectures preserve the mobility of different body parts, enabling controlled fusion between streams. Although such streams remain independent at the motion sources, they often require longer processing times and may lack explicit linguistic and semantic information. The MediaPipe Hands framework is used for real-time gesture and hand tracking, providing a lightweight and robust solution under varying illumination or different working environments [28]. The main factors that decrease performance over time include partial occlusions (objects from the background/working environment), gestures or hand overlap, blurry motions and jitter from irregular movements. For dynamic gestures, these factors are amplified, leading to discontinuities in hand landmark trajectories, particularly during rapid transitions between gestures or sign sequences. To obtain sustainable sequences, preprocessing strategies are applied, such as coordinate normalization (from absolute to relative), rotation, and gray scale conversion, which are suitable for models that learn from temporal dependencies [28]. While recent Transformer-based systems demonstrate strong temporal reasoning abilities, most operate on fused hand representations without separation, limiting robustness in asymmetric dynamic gestures [29,30].

The proposed two-handed dynamic recognition framework addresses three key challenges: first, the limitation of existing models in handling asymmetric and asynchronous hand trajectories; second, the ability of the Transformer-based cross-hand fusion mechanisms to improve temporal learning during the early recognition phase; and third, maintaining computational efficiency while modeling bidirectional spatial and temporal dependencies in real-time and real-world working environments. To address these challenges, the proposed dual-stream recognition system will process each hand independently and learn through the attention-based fusion encoder the cross-hand dependencies, aligning the misaligned spatial and temporal motion patterns. Despite significant progress in dynamic gesture recognition, two-handed sign language recognition still faces complex limitations in real-world scenarios. Requirements include a large dataset highly influenced by illumination, high contrast, inter-signer variability and background occlusions [31].

Sequence-based models such as LSTMs and GRUs effectively preserve sequential information on a short time basis, but struggle with long-range gestures and hand trajectories. Although the Transformers have a strong temporal reasoning performance, they remain strictly dependent on cross-hand (both hands fuse into one single representation). This design motivates the proposed dual-stream architecture, which explicitly models each hand independently before attention-based cross-hand fusion. The following sections present the

data acquisition phase, the preprocessing pipeline, architectural design and also the training strategy of the proposed dual-stream BiLSTM–Transformer dynamic recognition system.

3. Proposed Dual-Stream Recognition System

This section presents the architectural design of the proposed real-time two-handed dynamic gesture recognition system based on the BiLSTM–Transformer model, including dataset acquisitions, hand keypoints representation, landmark extraction and preprocessing using the MediaPipe Hands framework, and data augmentation. The overall architecture and data flow for the proposed system are summarized in Figure 1.

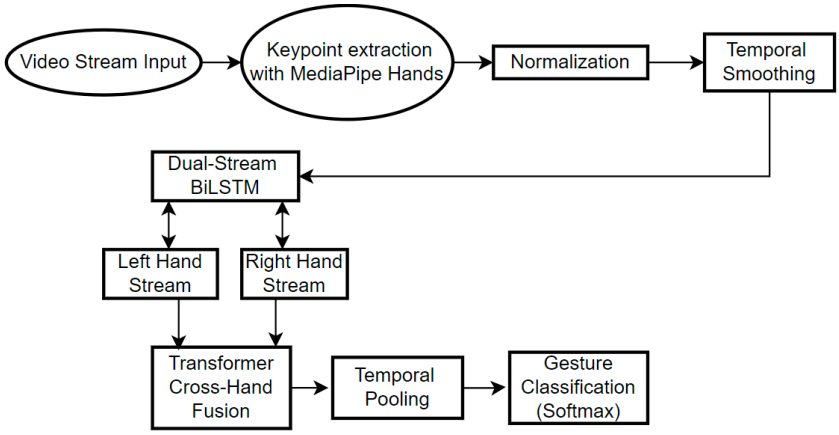

Figure 1. The AI processing pipeline of the proposed dual-stream BiLSTM–Transformer system.

The pipeline highlights the separation between preprocessing, dual-stream temporal encoding and attention-based cross-hand fusion mechanisms prior to final gesture classification.

3.1. Datasets Acquisition

To evaluate performance on real-world scenarios, the proposed dynamic recognition system used two dedicated datasets collected from the Romanian Sign Language alphabet. The complete dataset contains ten different dynamic class gestures, including both symmetric and asymmetric motion patterns. For each class, a fixed number of 100 temporal sequences per gesture were used to identify, label and train the model. This setup ensures consistent data representation and improves the stability of temporal learning. Data augmentation techniques were applied to increase the number of training samples and improve robustness under inter-signer variability in real working environments. To improve the diversity of gesture trajectories and increase robustness to execution variability, several augmentation operations were applied directly to hand landmark sequences. Each generated gesture sequence has two additional augmented variants. The augmented pipeline included the following operations: Gaussian coordinate jitter was applied to hand landmark coordinates with a zero mean to standard deviation σ of 0.001; global spatial factor randomly sampled from the interval 0.96–1.04; in-plane rotation around the wrists with angle sampled from the range of -6° to $+6^{\circ}$; and temporal warping through linear interpolation by modifying sequence speed with a factor sampled from 0.10 to 1.0, followed by resampling to the fixed sequence length of 30 frames. These operations preserve gesture semantics, increasing variability in execution speed, trajectory amplitude and signer execution style.

After augmentation, the final dataset contained approximately 3000 augmented samples (300 per gesture class) with preserved semantic consistency. Gesture recordings were

collected indoors using a standard RGB camera under mixed natural and artificial lighting conditions with unconstrained background environments.

All sequences were temporally fixed to 30 FPS to ensure stable modeling. Shorter sequences were padded and downsampled to preserve hand trajectory dynamics and maintain input dimensionality. This procedure ensures reproducibility across all gesture classes. Each gesture is represented as a numerical time series derived from video frames using the MediaPipe Hands framework. For each frame, a total of 42 hand landmarks are extracted, represented by normalized coordinates from absolute to relative (tridimensional x, y, and z). For each step, a 126-dimensional feature is encoded as a vector (2 hands $\times$ 21 landmarks $\times$ tridimensional coordinates). The final model input consists of fixed-length temporal sequences of the hand landmark vectors. For evaluation, the proposed dataset was split into three sets: training, validation, and testing subclasses, to prevent data leakage and ensure that samples from the original dataset were not distributed to other subsets. The proposed dataset focuses only on isolated two-handed dynamic gestures, which represent an important step toward robust modeling of the hand trajectories (asymmetric and asynchronous). While larger-scale datasets with diverse signer demographics are desirable, the proposed dataset was designed to prioritize controlled gesture execution and temporal consistency. A continuous sign language recognition system introduces supplementary challenges, such as gesture boundary ambiguity and coarticulation effects, which are considered explicitly as a future research direction. Due to privacy and data protection constraints, visual examples of signers performing dynamic gestures cannot be publicly shared. All recordings were anonymized and used exclusively for numerical hand landmark extraction, ensuring that no identifiable visual information is retained or distributed.

3.2. MediaPipe Hands Keypoint and Landmark Extraction

The proposed system uses the MediaPipe Hands framework to detect and preprocess the hand landmark keypoints under different working environments. For each hand, 21 degrees of freedom are extracted for each frame, including wrists and finger joints, where absolute coordinates are converted into relative values matching the input dimensionality. This transformation reduces sensitivity to background variations and improves robustness to illumination changes, which represent critical factors in dynamic gesture recognition. For each frame, MediaPipe returns the detected hand landmarks coordinates for both hands, enabling robust bilateral hand tracking.

The framework processes sequences while consistently tracking both hands throughout gesture execution. To ensure temporal continuity, frames affected by hand trajectory loss are temporally interpolated, corrected, or abandoned, depending on sequence stability constraints. Several factors directly affect real-time inference, including hand overlap, fast gesture transitions and partial or complete occlusion. Also, these effects may lead to discontinuous sequences, introducing noise and jitter during the recognition phase. To stabilize the gesture sequences, additional preprocessing steps were applied to improve robustness under inter-signer variability, as described in the following subsection.

3.3. Preprocessing Techniques on Sequences of Gestures

As mentioned in the previous section, the preprocessing techniques were applied to ensure robustness under signer variability and independence from a specific working environment. These techniques primarily target noise reduction caused by hand tracking instability under varying scales and viewpoint changes. Preprocessing was applied independently for each hand to preserve the trajectory integrity and temporal semantic patterns. The first preprocessing step involves coordinate normalization by converting

landmarks into relative representations. This transformation enables global adaptation to hand translation, camera (viewpoint) and signer-independent hand geometry through scale normalization. Additionally, scale normalization improves invariance to camera position and user hand size.

The second processing step addresses temporal jitters and discontinuities across consecutive gesture sequences. Often, dynamic gestures tend to fluctuate between rapid transitions of frames with a high impact on the hand keypoints. To stabilize hand landmark trajectories and reduce frame-to-frame jitter, a temporal smoothing operation was applied independently to each coordinate trajectory (x, y, and z) for both hands. The smoothing procedure was made using a moving-average filter with a window size of three frames. This step reduces high-frequency noise induced by hand tracking, preserving natural motion dynamics from gesture sequences. After smoothing, the trajectories remain temporally consistent and suitable for subsequent sequence modeling. The third preprocessing step enforces a fixed temporal length F with a total of thirty frames to ensure consistent model input. For instance, if a sequence contains fewer than thirty frames, padding is applied to ensure motion continuity. On the other hand, sequences that exceed thirty frames are downsampled to preserve hand trajectories and gesture dynamics. The preprocessed sequences are organized into two parallel inputs corresponding to each hand trajectory. For each trajectory, a dual-stream sequence of gestures was represented by the hand landmarks used in the training of the proposed model (BiLSTM–Transformer). As mentioned earlier, the augmentation techniques were applied to enhance generalization across additional training samples. To improve clarity and reproducibility of the proposed framework, the overall recognition pipeline is summarized in Algorithm 1.

Algorithm 1: Dual-Stream Dynamic Gesture Recognition Pipeline

Input: Video sequences
Output: Predicted gesture class

1. Extract hand landmark for each frame using MediaPipe Hands
2. Normalize landmark coordinates and apply temporal smoothing
3. Construct left-hand and right-hand landmark sequences
4. Encode each trajectory using BiLSTM encoders
5. Concatenate encoded representations ($M = \left[M^L ; M^R \right]$)
6. Apply Transformer self-attention to model cross-hand dependencies
7. Perform temporal pooling on the fused representation (M)
8. Classify the gesture using a fully connected layer followed by Softmax activation

3.4. BiLSTM Dual Encoder

To avoid the limitation of early fusion and preserve hand semantics and dynamics, the proposed system processes both hands' trajectories independently. This design enables the model to learn distinctive motion patterns for symmetric or asymmetric dynamics.

The trajectory-based representation of each dynamic gesture sequence is defined as follows (1):

$$S^L = \left[S_1^L, S_2^L, \ldots, S_F^L \right], \; S^R = \left[S_1^R, S_2^R, \ldots, S_F^R \right] \quad (1)$$

where S^L, $S^R \in \mathbb{R}^{F \times 63}$ denote the feature sequences for the left and right hand and F represents the number of frames (set to thirty). Each frame contains 21 hand landmarks described by tridimensional coordinates (x, y, and z), resulting in a 63-dimensional feature representation per hand. Each dual stream is encoded using BiLSTM to model bidirectional temporal dependencies along with gesture trajectories (see Equation (1)). This encoding

enriches contextual and semantic information, reducing confusion between gestures with similar motion patterns. Also, two independent BiLSTM encoders are employed for each hand using the following representations (2):

$$E^L \in \mathbb{R}^{F\times H},\ E^R \in \mathbb{R}^{F\times H} \tag{2}$$

where E^L and E^R denote the spatiotemporal BiLSTM encoders for both hands trajectories and H denotes the dimensionality for each representation. The encoded representations are obtained through two independent BiLSTM encoders applied to the landmark sequence of each hand using the following Formulas (3) and (4):

$$E^L = BiLSTM\left(S^L\right) \tag{3}$$

$$E^R = BiLSTM\left(S^R\right) \tag{4}$$

The BiLSTM layers capture both forward and backward temporal dependencies along the gesture trajectory. These representations enable bidirectional temporal modeling of hand trajectories and serve as input for the Transformer-based cross-hand fusion stage described in the following section [4].

3.5. Transformer Fusion Encoder

The Romanian Sign Language alphabet relies on hand semantics, motion trajectories and signed movement patterns. As mentioned earlier, early fusion approaches concatenate both hands into one single representation, forcing synchronicity and suppressing asymmetric gestures. This limitation can be addressed using a Transformer encoder with cross-hand fusion. The encoder learns semantic dependencies from both hands in a bidirectional and bilateral manner through self-attention mechanisms. The temporal fusion phase operates independently of BILSTM encoders, leveraging the high representation of the hands' trajectories rather than merging both into a single representation.

This self-attention fusion representation is defined by formula (5):

$$M = \left[\, M^L;\, M^R \right] \tag{5}$$

where M^L and M^R denote the temporal feature representations of the left- and right-hand trajectories obtained independently by BiLSTM encoders. These representations are concatenated to form the joint feature matrix ($M = \left[\, M^L\, ;\, M^R \right]$), which is then used as the input to the Transformer self-attention mechanism for modeling cross-hand dependencies. The concatenation is performed along the feature dimension, resulting in a fused representation $M \in \mathbb{R}^{F\times 2H}$, where F denotes the temporal length of the gesture sequence and H represents the dimensionality of the encoded BiLSTM features. The interaction between the two hand representations is then modeled through the Transformer self-attention mechanism defined in Formula (6):

$$Attention\left(\, M^L,\, M^R\right) = Softmax\left(\, M^L\left(\, M^R\right)^T\right) \tag{6}$$

This fusion strategy enables the system to learn from both hands without suppressing input trajectory information. The cross-hand attention module enables the model to dynamically learn the relative semantic contribution of each hand. Rather than enforcing temporal synchronization, the module assigns adaptive weights to the left- and right-hand representations based on the gestures' specific motion pattern. This mechanism allows the dominant or complementary hand to emerge naturally during the training process. The

concatenated representation M is refined through the Transformer attention mechanism, and the resulting attention representation is further used in the temporal pooling stage described in Section 3.6 [5].

3.6. Temporal Pooling and Classification

The fused representation obtained from the BiLSTM and Transformer module is converted into a compacted version suitable for gesture classification. For each dynamic gesture sequence, the objective is to predict a single class label. To achieve this, temporal pooling is applied to eliminate length dependency in the fused representation. The fused representation obtained from the Transformer cross-hand attention module is denoted as (M).

This representation results from concatenation of the encoded hand trajectories M^L and M^R, followed by the Transformer self-attention mechanism that models cross-hand dependencies. Temporal pooling is then applied to the fused matrix (M) to obtain a global embedding vector (x), defined in Equation (7):

$$x = Pool(M) \tag{7}$$

This vector encodes the temporal evolution of each dynamic gesture learning from fused dependencies of the representation. Final gesture classification is performed using an MLP head (the fully connected layer), followed by the *Softmax* activation over the D gesture classes, represented in Equation (8):

$$f = Softmax(ax + b) \tag{8}$$

where a and b denote the trainable parameters of the classifier, x denotes pooled gesture embedding and f denotes the predicted class probabilities. In this study, the number of classes corresponds to ten gesture categories included in the datasets [32].

3.7. Training andHyperparameters

The model was trained in a supervised manner on an augmented dataset consisting of ten different classes of gestures. The Adam optimizer was employed to ensure fast convergence and stable temporal modeling. The sparse categorical cross-entropy loss was computed using the following formulation (9):

$$L = -\sum_{d=1}^{D} u_d \log(\hat{u}_d) \tag{9}$$

where u_d represents the ground truth label for the gesture class d, and $\hat{u}_d$ denotes the predicted output of the model. The batch size of the dataset was selected to enable stable parameter updates and reduce training bias across epochs. To prevent overfitting and improve generalization, regularization strategies were applied. Dropout regularization layers were applied within temporal modeling and fusion stages, along with early stopping based on validation loss. Additionally, a dynamic learning rate strategy was employed to improve convergence stability, while data augmentation further enhanced robustness under different real-world scenarios and inter-signer variability. The complete pipeline was realized entirely in Python 3.10 using deep learning libraries, and the final model was optimized for real-time operation on compact hand landmarks rather than raw RGB video sequences [33].

The BiLSTM encoders use 64 hidden units per direction, while the Transformer encoder consists of three layers with five attention-based heads per layer. The dropout rate applied was set to 0.2, and the model was optimized with Adam from an initial learning rate of 0.0001 on a batch size of 32 units. To maintain a lightweight design, the proposed architecture used approximately 1.6 million trainable parameters (reduced complexity and suitable for real-time deployment applications). The number of hidden units and the Transformer encoder layers were empirically selected as a representation trade-off between representational capacity and computation efficiency. Increasing model depth or hidden dimensionality resulted in higher inference latency, a critical component for real-time deployment, without yielding significant accuracy improvements.

3.8. Performance Evaluation Metrics

For the proposed recognition system, specific metrics (accuracy, precision, recall and F1-score) were used to quantify overall performance in the dynamic gesture recognition. These metrics were derived from the confusion matrix and computed as follows based on formulas (10)–(13):

$$Accuracy = \frac{T^+ + T^-}{T^+ + T^- + F^+ + F^-} \quad (10)$$

$$Precision = \frac{T^+}{T^+ + F^+} \quad (11)$$

$$Recall = \frac{T^+}{T^+ + F^-} \quad (12)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (13)$$

where T^+ denotes true positives (correctly predicted gesture samples of a given class), T^- denotes the true negative, F^+, denotes false positives, and F^- denotes false negatives [34].

3.9. Communication and Translation

The proposed system integrates not only dynamic gesture recognition, but also a bidirectional communication assistant enabling interaction with users who rely on sign language communication. Additionally, the system incorporates text generation and speech synthesis to provide real-time vocal feedback and improve stability. The proposed framework consists of four main components: dynamic two-handed gesture recognition from the RSL alphabet, translation from gesture to text, syntactic validation using the Stanza library and vocal TTS feedback in real time. The gesture recognition module predicts and labels each input sequence using the Dual-Stream BiLSTM–Transformer architecture. The resulting label is mapped to a word token and then added to the gesture-to-text vocabulary. Each word token is appended to a sentence buffer, enabling text-generating signing sequences. The proposed architecture, overall, integrates an assistive solution of dynamic gesture recognition with grammatical and vocal analysis for multimodal usage (see Figure 2) [35].

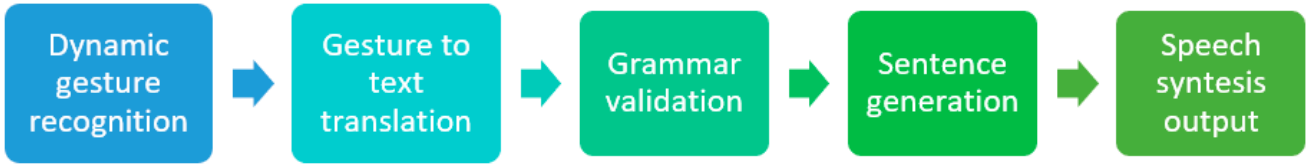

Figure 2. The hybrid architecture of the proposed system.

The generated text stream is parsed and syntactically analyzed through the Stanza library. In this phase, all sentences generated from gestures are validated and then grammatically corrected based on the structure estimation made independently by Stanza (supporting strategies include filtering invalid sequences and resolving inconsistencies in word order). In this way, the system gains usability, accessibility and high robustness levels in practical scenarios (in continuous streams from real-world communication scenarios) [35]. The text is then processed through a vocal feedback TTS, enabling signers to have auditory support. Speech synthesis enhances everyday usability, particularly in public environments where users receive immediate spoken interpretation of recognized gestures.

3.10. Experimental Setup and Evaluation Protocol

The proposed system was evaluated using the Romanian Sign Language gesture vocabulary dataset described in Section 3.1. Each gesture sequence was recorded at 30 FPS and later normalized to fixed-length sequences. The reference to minimum frame counts reflects the temporal length of gesture sequences rather than the acquisition frame rate. Generalization capability was enhanced through data augmentation techniques, increasing motion pattern diversity while preserving gesture semantics.

The dataset was split into training 70%, validation 15% and testing 15%. All experiments were conducted three times using different random initializations. The system was implemented using TensorFlow Lite, Keras and MediaPipe Hands for hand landmark extractions. Real-time evaluation was conducted on a standard server with an equipped CPU and a powerful integrated GPU; however, inference performance was measured on CPU-only execution [36]. All experiments were conducted on a workstation equipped with an Intel i7-class processor, 16 GB RAM and integrated graphics. Inference performance was evaluated using CPU-only execution to simulate real-world deployment conditions [36].

4. Results

This section presents the experimental validation results of the proposed two-handed dual-stream BiLSTM–Transformer gesture recognition architecture. The evaluation focuses on the confusion matrix, robustness under different working environments and the system's performance within an assistive bidirectional communication framework [37].

4.1. Recorded Overall Performance

The proposed architecture achieved an overall recognition accuracy of 96% on the testing dataset. This result indicates effective independent modeling of each hand through the attention mechanisms. Based on the precision and recall values, the system does not favor a specific gesture class and maintains stable performance across the entire vocabulary. The standard evaluation metrics provided a balanced overview of the recognition performance across all gesture classes. For most gesture classes, recognition performance was close to perfect, with slightly low recall values observed for certain gesture classes such as Request and Appreciation. These variations are attributed to differences in signer execution style and temporal motion dynamics. Overall, the evaluation metrics indicate strong performance and generalization capability, with minimal confusion between dynamic gesture sequences [38].

A detailed class-level performance report on the independent testing dataset, including precision, recall, F1-score and support values is presented in Table 1.

Table 1. The gesture class performance of the proposed dynamic recognition system.

Gesture Class	Precision	Recall	F1-Score	Samples Support
Hello	90	95	93	300
Thank you	95	100	98	300
I love you	95	100	98	300
You	100	100	100	300
Me	95	100	98	300
Request	86	90	88	300
Yes	100	100	100	300
No	95	90	92	300
Apology	100	95	97	300
Appreciation	100	85	92	300

4.2. Analysis of the Confusion Matrix

The confusion matrix in Figure 3 indicates strong classification performance reflected by the dominant values along the main diagonal, where most gesture samples are correctly classified.

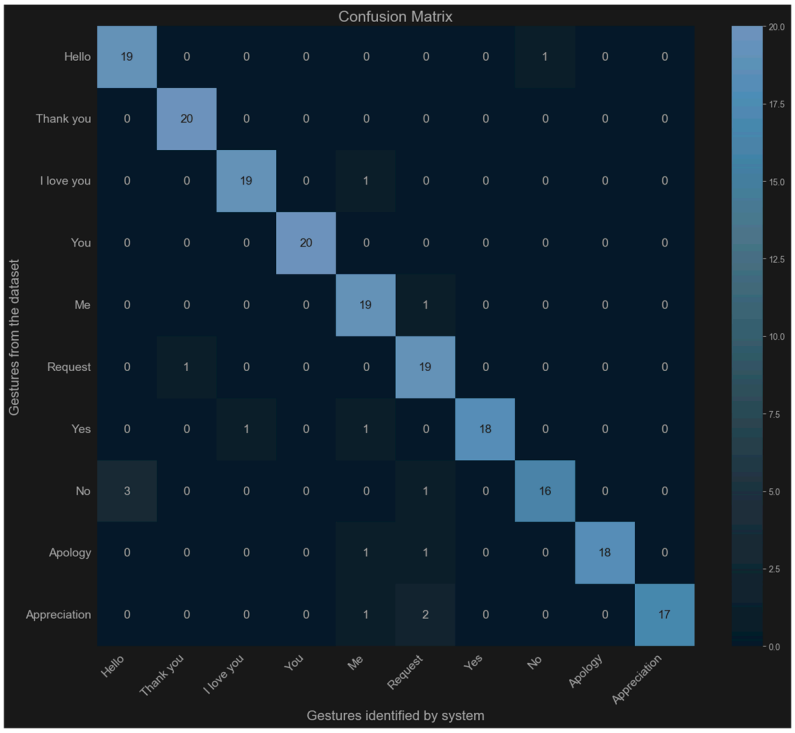

Figure 3. The confusion matrix of the proposed dynamic hybrid system.

Minor misclassification was observed among visually similar gesture classes that share comparable motion directions or overlapping execution phases. The limited confusion supports the effectiveness of the proposed dual-streamed architecture, preserving hand semantics independently and reducing gesture ambiguity observed in early-fusion approaches. Motion blur, as well as partial occlusions, may affect the extracted temporal parameters during gesture execution. Another contributing factor is inter-signer variability, particularly differences in gesture execution speed and transition dynamics.

4.3. Ablation Study and Architectural Impact Analysis

To assess the contribution of each architectural component, an ablation study was conducted on the proposed recognition system. Several variables were studied, such as a

single-streamed BiLSTMs model concatenated with keypoints of the hands, a dual-stream model without Transformer fusion and a merged model with unified trajectories. The results indicate that merging trajectories from both hands significantly decreases robustness in pattern recognition, ignoring asynchronous or asymmetric gesture classes. In this case, temporal inconsistencies between both hands affect the network's capability to retain information regarding gesture execution and the completion phases, especially when long-range sequences of gestures are not modeled using the Transformer module. The quantitative results of the ablation study are summarized in Table 2. The results indicate that the proposed system performs better when the left- and right-hand streams are processed independently compared to a single-stream BiLSTM. The inclusion of the cross-hand fusion module provides an additional performance gain, confirming the effectiveness of the proposed architecture.

Table 2. Ablation study evaluating individual architectural components.

Model	Architecture Representation	Precision (%)	Recall	F1-Score
Single-stream BiLSTM	Unified representation without separation	88	86	86
Dual-stream BiLSTM	Independent temporal modeling for left/right hand	93	91	91
Dual-stream BiLSTM + Transformer	Cross-hand attention-based fusion	96	95	95

The complete architecture achieves the best performance across all evaluated gesture classes. The Transformer-based fusion mechanism enables adaptive weighting of both hands, dynamically focusing on the dominant or complementary hand while integrating motion cues from gesture patterns. This plays a critical role in attention-based modeling, enhancing robustness under different working conditions and gesture speed execution. The ablation experiments demonstrate that the proposed recognition system achieves high accuracy while maintaining computational efficiency suitable for real-time deployment [39].

4.4. Real-Time Deployment Performance Considerations

The complete recognition system also targets real-time deployment in assistive real-world scenarios of communications without reliance on high-end CPU/GPU hardware. The framework, based on MediaPipe Hands, processes hand landmark keypoints on a CPU-only server to assess deployment feasibility. MediaPipe has demonstrated stable and efficient real-time tracking under moderate conditions, including varying illumination, complex background, shadows and partial occlusions) hand landmarks, and semantics. Computational complexity was significantly reduced by operating on compact hand landmark characteristic vectors rather than raw RGB frames, leading to improved efficiency compared to CNN approaches. A significant improvement is achieved by processing compact dual-hand sequences in parallel [40].

The fusion module embeds temporal and high-resolution feature representations while maintaining real-time inference, although limited by the computational cost of attention-based mechanisms. As a result, the system maintains stable performance from a deployment perspective, enabling practical assistive communication for users who utilize sign language. The CPU-only configuration removes dependency on specialized hardware, making the system viable for real-time dynamic RSL interpretation. The average end-to-end inference latency was approximately 21–30 ms per frame, while the model-only time was around 15–18 ms on a CPU-only system (Intel i7 class processor). These results confirm that the proposed dual-stream architecture maintains real-time performance while preserving recognition accuracy, making it suitable for real-time deployment on resource-constrained devices.

4.5. Summary of Experimental Results

The experimental results from the proposed system provide strong improvements in dynamic dual-streamed gesture recognition, particularly for misaligned sequences. By preserving semantic information within spatial–temporal representations, the proposed system effectively discriminates between motion patterns of each hand independently of execution speed or gesture style, outperforming existing approaches. The ablation study confirms that the architecture provides a meaningful contribution, maintaining low computational cost under CPU-only deployment and supporting real-time applications. Overall, the findings establish a balance trade-off between recognition accuracy, model adaptability and real-world sign language interpretation for practical applications. A detailed discussion of the results, limitations and future implications is provided in the next section.

4.6. Comparison with Recent Approaches

This section presents a comparison with recent approaches that rely on gesture recognition from the sign language alphabet. The systems are grouped into hand landmark models and Transformer-based and deep learning video-based approaches. Video-based approaches include CNNs and Transformer models that learn features from RGB frame sequences as presented in [6,9,27]. Although these approaches achieve high performance, they require substantial computational resources due to the processing of full-resolution image sequences. This represents a major limitation for real-time inference from CPU-based systems. In contrast, the proposed system operates on MediaPipe hand landmarks, reducing input dimensionality while preserving essential motion information.

Hand landmark RNN models have demonstrated strong temporal modeling performance using LSTM or GRU fusion strategies, as reported in [8,34,36]. However, such designs may suppress asymmetric hand dynamics, potentially leading to reduced robustness. To address this limitation, the proposed study preserves complex dynamic semantics by modeling both hands independently, improving accuracy and robustness. Transformer-based self-attention mechanisms are capable of modeling long-range dependencies in gesture sequences, as shown in [4,14,22].

These approaches typically operate on fused inputs or focus primarily on continuous sign language recognition. The proposed framework establishes explicit interactions between BiLSTM encoders through cross-hand mechanisms. This enables dynamic adjustments over time, enhancing complementary gestures and asymmetric roles. Compared to the sensor or glove-based systems, such as in [13,20], the proposed method achieves competitive performance with lightweight computation requirements, supporting practical deployment.

Table 3 presents a comparison of recent approaches in terms of hand modeling strategy, fusion type, target focus, number of hands processed and deployment requirements. Many of these approaches rely on specialized hardware (wearable watches/sensors, colored gloves, and radar wavelet sensors) or require substantial computational resources. Despite these limitations, the proposed system provides a scalable solution that models each hand independently within a dual-stream architecture prior to attention-based mechanisms. This design enables real-time performance for asynchronous two-handed dynamic gestures, which are less addressed in many recent approaches.

Table 3. A comparison of the existing gesture recognition approaches.

Reference	Data Input	Main Focus	Real-Time Inference (CPU)	Hand Recognition	Architecture	Fusion Type
[6]	RGB video	Continuous SRL	No	One hand	Transformer	Sensor
[4]	Wearable sensors	Dynamic SLR	No	Sensors	Video + Transformer	Early
[8]	Skeletal landmarks	SLR	No	Two hands	CNN/LSTM	Late
[13]	Optical glove sensors	SLR	Yes	Wearable sensor	MLP	Sensor
[14]	RGB + pose	Continuous SRL	No	Two hands	Transformer	Multimodal
[20]	Wave radar	Dynamic	Yes	Radar sensors	CNN/RNN	Sensor
[34]	Skeletal graphs	Dynamic	No	Two hands	GCN	Graph
[36]	Hand landmarks	Dynamic	Limited	Single stream	CNN/LSTM	Early
Proposed System	MediaPipe Hands landmarks	Two-hand Dynamic RSL	Yes	Dual-stream	BiLSTM + Transformer	Cross-hand

5. Discussion

The obtained results highlight the importance of independently modeling both hands. Unlike existing approaches, the proposed recognition system preserves distinctive spatial–temporal patterns from both hands. This design introduces a beneficial motion component that enables learning from dynamic patterns without trajectory overlap, thereby improving robustness. The integration of self-attention-based Transformer mechanisms brings an additional layer that aligns contextual information from misaligned motion sequences. The ablation study demonstrates that performance improves significantly when the Transformer model is combined with BiLSTM encoders. Using skeletal hand landmarks instead of raw video sequences reduces computational complexity and storage requirements while preserving essential motion information. This approach supports real-time inference and improves generalization across different working environments (illumination changes, occlusion, and shadows), remaining suitable for deployment on constrained resource devices. Despite the promising results, several limitations must be considered. The proposed study focuses on isolated gestures with a limited number of classes, which may restrict generalization across broader signer populations within Romania. Continuous sign language recognition involves unconstrained gesture transitions and contextual dependencies, which are not addressed in the present research, but will be investigated in the future through online temporal segmentation and over larger-scale datasets. Although the evaluated gesture vocabulary is limited, the objective of this study is to analyze asymmetric and asynchronous two-handed motion modeling under controlled conditions rather than to provide comprehensive sign language coverage.

Another key finding from an application perspective is the integration of the assistive bidirectional communication component. CPU-based deployment enables transition to portable devices, educational environments, public services and everyday assistive applications for individuals who use sign language. Therefore, the system contributes not only methodologically, but also practically by bridging accessibility and communication gaps in human–computer interaction [36,37].

Results Discussion

Despite promising results, several limitations of the proposed dynamic recognition system should be acknowledged. The dataset used contains a limited number of dynamic gesture classes and signers. Although data augmentation techniques were applied to

improve variability, the model may remain sensitive to very fast motion speeds and to new, unseen or complex gestures. The absence of explicit linguistic modeling represents another limitation, even though hand landmark representations are computationally efficient. Fine-grained visual cues, such as detailed hand shape, size, face recognition, and emotional expressions, could further improve the semantic relevance of sign language recognition.

Future multimodal research will explore advanced fusion strategies that incorporate facial recognition features to enrich semantic and contextual understanding [26,28,29]. Another important limitation relates to the absence of temporal segmentation. The proposed system operates on fixed-length sequences, which simplifies training, but does not fully reflect continuous signing without explicit boundaries. Implementing online segmentation for continuous sign language streams would significantly improve communication with individuals who use sign language.

Finally, although the proposed model is lightweight for CPU-based inference, latency may increase over time when multiple tasks are executed simultaneously. To further improve the system, model compression techniques, such as pruning or quantization, could facilitate efficient edge deployment [32,33].

6. Conclusions

The proposed system extends conventional dynamic gesture recognition for the RSL alphabet by integrating bidirectional communication, speech synthesis and syntactic validation feedback. By integrating real-time recognition, the proposed architecture contributes to bridging the communication gap for individuals who use sign language through a natural vocal interaction. The dual-stream BiLSTM–Transformer recognition system enables accurate modeling of two-handed signs and remains computationally efficient for real-world deployment on CPU-based systems. This approach provides a balanced and scalable solution that can be integrated into educational tools and public service applications to support social inclusion and human–computer interaction.

Moreover, the modular architecture allows easy expansion of the gesture vocabulary and inclusion of additional gesture classes with intra-class variations supporting improved semantic representation. Future work will focus on expanding the dataset, including new sign language alphabets, to further enhance scalability and real-world assistive social inclusion.

Author Contributions: E.A., the corresponding author was responsible for the conception of the proposed study, the experimental implementation, the dataset analysis, and the preparation of the original manuscript. T.C.-O. provided the scientific supervision, methodological validation, and critical evaluation of the proposed architecture. C.G. contributed to the refinement of the methodological approach and the revision of the manuscript. A.D.-A., U.A.-G., and S.B.-C. participated in the data acquisition phase, experimental real-world validation, and analysis of the obtained and final results. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The main contributions presented in the study are included in the current research; further inquiries can be addressed directly to the corresponding author. The generated dataset during this study consists of a custom RSL dynamic gestures alphabet. Due to privacy and informed consent restrictions, raw video recordings cannot be publicly released.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LSTM	Long-Short-Term Memory
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GCNs	Graph Convolutional Networks
GRUs	Gated Recurrent Units
MLP	Multilayer Perceptron
BiLSTM	Bidirectional Long-Short-Term Memory
RSL	Romanian Sign Language
RGB	Red Green Blue color model
GPU	Graphical Processing Unit
CPU	Central Processing Unit
FPS	Frames Per Second
TTS	Text-To-Speech

References

1. Salazar-Chacón, G.; Copa, M.E.J.; Caiza, D.M.; Aldas, C.P.; Guanga, C.G. Proof-of-Concept Translator Using AI: Ecuadorian Sign Language to Human Voice App. In Proceedings of the 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), Vellore, India, 22–23 February 2024; pp. 1–7. [CrossRef]
2. Thorat, N.N.; Kulal, N.M.; Baburao, K.V.; Hirve, S.A.; Kolekar, S.; Sangore, R. AI based Security in computer vision and machine learning algorithms for hand sign recognition. In Proceedings of the 2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), Pune, India, 17–19 October 2024; pp. 1–5. [CrossRef]
3. Zhang, H.; Wang, L.; Pei, J.; Lyu, F.; Li, M.; Liu, C. RF-Sign: Position-Independent Sign Language Recognition Using Passive RFID Tags. *IEEE Internet Things J.* **2024**, *11*, 9056–9071. [CrossRef]
4. Garg, M.; Ghosh, D.; Pradhan, P.M. MVTN: A Multiscale Video Transformer Network for Hand Gesture Recognition. *arXiv* **2024**, arXiv:2409.03890. [CrossRef]
5. Xu, F.; Chaudhary, L.; Dong, L.; Setlur, S.; Govindaraju, V.; Nwogu, I. A Comparative Study of Video-Based Human Representations for American Sign Language Alphabet Generation. In Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), Istanbul, Turkiye, 27–31 May 2024; pp. 1–6. [CrossRef]
6. Wang, Q.; Zheng, Z.; Wang, Q.; Deng, D.; Zhang, J. Generalizations of Wearable Device Placements and Sentences in Sign Language Recognition with Transformer-Based Model. *IEEE Trans. Mob. Comput.* **2024**, *23*, 10046–10059. [CrossRef]
7. Keerthana, M.; Kusuma, C.D.; Banakar, L.S.; Mallika, K. Transforming Model-to-Model Interactions with Advanced Natural Language Processing: A Deep Learning Approach. In Proceedings of the 2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET), Pune, India, 16–17 January 2025; pp. 1–6. [CrossRef]
8. Zhang, M.; Gao, Q.; Ju, Z. DHF-SLR: Dual-Hand Multi-Stream Fusion Network for Skeleton-Based Sign Language Recognition. In Proceedings of the 2024 International Conference on Advanced Robotics and Mechatronics (ICARM), Tokyo, Japan, 8–10 July 2024; pp. 649–654. [CrossRef]
9. Jim, A.A.J.; Rafi, I.; Tiang, J.J.; Biswas, U.; Nahid, A.-A. KUNet-An Optimized AI Based Bengali Sign Language Translator for Hearing Impaired and Non Verbal People. *IEEE Access* **2024**, *12*, 155052–155063. [CrossRef]
10. Punekar, R.; Borole, Y.; Futane, P. Enabling Communication Strategies: Indian Alpha-Numeric Sign Gesture Recognition. In Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 5–7 April 2024; pp. 1–6. [CrossRef]
11. Wyawahare, M.; Adbale, T.; Jorvekar, V. ListenBot: Augmented Reality Based Speech to Sign Language Conversion. In Proceedings of the 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 18–19 January 2024; pp. 332–339. [CrossRef]
12. Srivastava, J.; Madhava, V.S.; Karthikeya, P.R.; Reddy, M.S.; Koundal, S. Sign2Text Conversion. In Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024; pp. 1–5. [CrossRef]
13. Pal, D.; Kumar, A.; Kumar, V.; Basangar, S.; Tomar, P. Development of an OTDR-Based Hand Glove Optical Sensor for Sign Language Prediction. *IEEE Sens. J.* **2024**, *24*, 2807–2814. [CrossRef]
14. Hu, H.; Pu, J.; Zhou, W.; Fang, H.; Li, H. Prior-Aware Cross Modality Augmentation Learning for Continuous Sign Language Recognition. *IEEE Trans. Multimed.* **2024**, *26*, 593–606. [CrossRef]

15. Schuessler, C.; Zhang, W.; Bräunig, J.; Hoffmann, M.; Stelzig, M.; Vossiek, M. Radar-Based Recognition of Static Hand Gestures in American Sign Language. In Proceedings of the 2024 IEEE Radar Conference (RadarConf24), Denver, CO, USA, 6–10 May 2024; pp. 1–6. [CrossRef]
16. Schechter, C.R.; Gomes, J.G.R.C. Enhancing Gesture Recognition Performance Using Optimized Event-Based Data Sample Lengths and Crops. In Proceedings of the 2024 IEEE 15th Latin America Symposium on Circuits and Systems (LASCAS), Punta del Este, Uruguay, 27 February 2024–1 March 2024; pp. 1–5. [CrossRef]
17. Vijitkunsawat, W.; Chantagram, P. Deep Multimodal-Based Recognition of Thai Finger Spelling with Two-Handed Postures. In Proceedings of the 2024 12th International Electrical Engineering Congress (iEECON), Pattaya, Thailand, 6–8 March 2024; pp. 1–6. [CrossRef]
18. Zhu, R.; Fan, D.; Lin, J.; Feng, H.; Wang, H.; Dai, J.S. Machine-Learning-Assisted Soft Fiber Optic Glove System for Sign Language Recognition. *IEEE Robot. Autom. Lett.* **2024**, *9*, 1540–1547. [CrossRef]
19. Vijitkunsawat, W.; Anunvrapong, P.; Chantngarm, P. Human Joint Coordinate Sequencing in Video-Based Thai Finger Spelling Recognition. In Proceedings of the 2024 8th International Conference on Information Technology (InCIT), Chonburi, Thailand, 14–15 November 2024; pp. 681–686. [CrossRef]
20. Jin, B.; Ma, X.; Hu, B.; Zhang, Z.; Lian, Z.; Wang, B. Gesture-mmWAVE: Compact and Accurate Millimeter-Wave Radar-Based Dynamic Gesture Recognition for Embedded Devices. *IEEE Trans. Hum.-Mach. Syst.* **2024**, *54*, 337–347. [CrossRef]
21. Bersoto, J.C.P.; Centeno, C.J.; De Jesus, C.C.; Mangahas, J.C.; Pascual, E.S.; Sison, A.A.R.C. TalkEasy: An Interactive Web-Based Platform for Hiragana Japanese Sign Language Fingerspelling Acquisition using K-Nearest Neighbors (KNN) Algorithm. In Proceedings of the 2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS), İstanbul, Turkiye, 6–7 December 2024; pp. 1–6. [CrossRef]
22. Li, J.; Xu, J.; Liu, Y.; Xu, W.; Li, Z. Enhancing the Applicability of Sign Language Translation. *IEEE Trans. Mob. Comput.* **2024**, *23*, 8634–8648. [CrossRef]
23. Begum, H.; Chowdhury, O.; Hridoy, S.R.; Islam, M.M. AI-Based Sensory Glove System to Recognize Bengali Sign Language (BaSL). *IEEE Access* **2024**, *12*, 145003–145017. [CrossRef]
24. Padmakala, S.; Husain, S.O.; Poornima, E.; Dutta, P.; Soni, M. Hyperparameter Tuning of Deep Convolutional Neural Network for Hand Gesture Recognition. In Proceedings of the 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), Bengaluru, India, 9–10 August 2024; pp. 1–4. [CrossRef]
25. Singh, G.; RawatRawat, R.S.S.; Agarwal, V.; Manwal, M.; Purohit, K.C. Eloquence in Silence: A Boon for Speech and Hearing Impaired Community. In Proceedings of the 2024 International Conference on Computer, Electronics, Electrical Engineering & Their Applications (IC2E3), Srinagar Garhwal, India, 6–7 June 2024; pp. 1–6. [CrossRef]
26. Jeon, S.; Park, S.; Bae, J.; Lim, S. Applying Multistep Classification Techniques with Pre-Classification to Recognize Static and Dynamic Hand Gestures Using a Soft Sensor-Embedded Glove. *IEEE Sens. J.* **2024**, *24*, 30668–30679. [CrossRef]
27. Ramachandran, S.; Shreedar, R.M.; Narayana, K.E. Online Dynamic Hand Gesture Recognition Using 3D-CNN-RNN Hybrid Architecture. In Proceedings of the 2024 2nd International Conference on Networking and Communications (ICNWC), Chennai, India, 2–4 April 2024; pp. 1–6. [CrossRef]
28. Troconis, L.G.; Felicetti, R.; Monteriù, A. A Cutting-Edge Approach to Decision-Making in Awake Neurosurgery Based on Hand Motion Recognition. *IEEE Access* **2024**, *12*, 100760–100771. [CrossRef]
29. Gupta, S.; Bujade, P.; Singh, V.; Tiwari, D.S. Hand Gesture Recognition System Using Deep Learning. *Int. J. Multidiscip. Res.* **2024**, *6*. [CrossRef]
30. Cotteret, M.; Greatorex, H.; Ziegler, M.; Chicca, E. Vector Symbolic Finite State Machines in Attractor Neural Networks. *Neural Comput.* **2024**, *36*, 549–595. [CrossRef] [PubMed]
31. Rainio, O.; Teuho, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. [CrossRef] [PubMed]
32. Carvalho, M.; Pinho, A.J.; Brás, S. Resampling approaches to handle class imbalance: A review from a data perspective. *J. Big Data* **2025**, *12*, 71. [CrossRef]
33. Charan, S.S.; Srihitha, V.; Palaniswamy, S.; Chethana, S. Real-Time Speech-to-Text Holographic Communication for the Deaf Children and Elderly. In Proceedings of the 2024 Fourth International Conference on Multimedia Processing, Communication & Information Technology (MPCIT), Shivamogga, India, 13–14 December 2024; pp. 327–331. [CrossRef]
34. Naz, N.; Sajid, H.; Ali, S.; Hasan, O.; Ehsan, M.K. MSE-GCN: A Multiscale Spatiotemporal Feature Aggregation Enhanced Efficient Graph Convolutional Network for Dynamic Sign Language Recognition. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *9*, 2979–2994. [CrossRef]
35. Rastogi, R.; Arya, A.; Chaudhary, A.; Chola, A.; Josan, P.K. Study of Enhanced Communication of Human-Computer Using Gesture Recognition Technique. In Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 14–15 March 2024; pp. 2131–2136. [CrossRef]

36. Hax, D.R.T.; Penava, P.; Krodel, S.; Razova, L.; Buettner, R. A Novel Hybrid Deep Learning Architecture for Dynamic Hand Gesture Recognition. *IEEE Access* **2024**, *12*, 28761–28774. [CrossRef]
37. Liu, X.; Tang, S.; Zhang, B.; Wu, J.; Ma, X.; Wang, J. WiVi-GR: Wireless-Visual Joint Representation-Based Accurate Gesture Recognition. *IEEE Internet Things J.* **2024**, *11*, 2701–2711. [CrossRef]
38. Kothadiya, D.R.; Bhatt, C.M.; Saba, T.; Rehman, A.; Bahaj, S.A. SIGNFORMER: DeepVision Transformer for Sign Language Recognition. *IEEE Access* **2023**, *11*, 4730–4739. [CrossRef]
39. Coster, D.; Dambre, J. Sign Language Recognition with Transformer Networks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*; Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., et al., Eds.; European Language Resources Association: Marseille, France, 2020; pp. 6018–6024.
40. Camgoz, N.C.; Koller, O.; Hadfield, S.; Bowden, R. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10020–10030. [CrossRef]