International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-13, Issue-3, May 2025
https:/doi.org/10.55524/ijircst.2025.13.3.6
Article ID IRP-1629, Pages 35-46
www.ijircst.org

Real-Time Vision-Based Indian Sign Language Translation Using Deep Learning Techniques

Subham Pandey[1], Sumaiya Tahseen[2], Rohit Pathak[3], Hina Parveen[4], and Maruti Maurya[5]

[1, 2, 3] B. Tech Scholar, Department of Computer Science and Engineering, Integral University, Lucknow, India
[4, 5] Assistant Professor, Department of Computer Science and Engineering, Integral University, Lucknow, India

Correspondence should be addressed to Subham Pandey subham.integral@gmail.com

Received 3 April 2025; Revised 16 April 2025; Accepted 1 May 2025

ABSTRACT- This work proposes a vision-based approach to real-time sign language translation for Indian Sign Language (ISL). The system uses state-of-the-art deep learning architectures such as CNN (Convolutional Neural Networks), LSTM (Long Short-Term Memory) networks, and Transformer-based encoder-decoder models for gesture recognition in both isolated and continuous forms. Data preprocessing techniques such as DTW (Dynamic Time Warping) were applied to augment and normalize gesture sequences from custom ISL and public ASL datasets. The model performance was quantitatively evaluated using precision, recall, F1-score, BLEU, ROUGE, CER(character error rate) and WER (word error rate).

A Transformer-based model outperformed the achieving a BLEU score of 0.74 and a classification accuracy of 96.1%. The developed desktop application enables real-time ISL-to-English translation at 18 FPS without requiring external sensors, while ablation studies validate the benefits of multimodal fusion and pose-language alignment. This work demonstrates a robust, scalable approach to non-intrusive sign language translation, advancing accessibility for the DHH community.

KEYWORDS-Transformer-based Encoder-Decoder, Spatiotemporal Gesture Modeling, Indian Sign Language (ISL), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Dynamic Time Warping (DTW), Real-time Sign Language Translation.

I. INTRODUCTION

Communication is an essential aspect of human life, yet millions of people who are hard of hearing or deaf experience significant challenges in this area around the world continue to experience barriers due to the limited adoption and understanding of sign languages in mainstream society [5], [12]. Sign languages, being natural and visually rich, vary significantly across regions, with no universal standard, making the creation of a robust translation system both necessary and challenging [6], [21], [28]. Despite being linguistically complete, sign languages remain underrepresented in technological solutions for accessible communication.

Recent advances in artificial intelligence (AI), computer vision, and deep learning have made the doors open for real-time sign language recognition and interpretation Vision-based techniques based on convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) models, and transformer-based architectures have achieved high accuracy in gesture classification and sequence learning [14], [22], [30], [38]. These models allow systems to learn the spatial-temporal patterns of hand movements and facial expressions are important in guaranteeing correct sign interpretation.

This research focuses on a vision-based AI model for sign language translation, particularly emphasizing Indian and American Sign Languages, without reliance on wearable sensors or external hardware [9], [16], [29]. The aim is to provide a non-intrusive, real-time system that captures visual inputs via standard cameras, extracts relevant features, classifies them into sign tokens, and then translates these into meaningful text or speech.

Furthermore, this paper addresses the social exclusion faced by the deaf community due to linguistic isolation and evaluates how AI systems can contribute to bridging this gap. The methodology involves camera-based image acquisition, pre-processing for background removal and contrast enhancement, feature extraction using CNNs, classification through hybrid models like CNN-SVM, and natural language generation via encoder-decoder models [13], [26], [31], [40]. It considers the non-manual markers like mouth patterns and facial expressions in enhancing translation accuracy.

By leveraging publicly available datasets and applying state-of-the-art AI models, this work proposes a scalable solution to reduce communication barriers and increase inclusivity for individuals with hearing and speech impairments.

II. BACKGROUND

Sign languages are fully-fledged natural languages that have evolved independently of spoken languages and exhibit all essential linguistic features such as grammar, morphology, and syntax [6], [21]. In contrast to spoken languages, sign languages adopt a visual-gestural modality, and meaning is drawn from hand shapes, movement, orientation, facial expression, and body

posture [28], [33]. There are more than 200 identified sign languages today, each formed by its geographical, cultural, and social backgrounds [25], [39].

Indian Sign Language (ISL), the primary focus of this study, has developed organically across various regions of India and remains under-documented compared to British Sign Language (BSL) or American Sign Language (ASL) [18], [26], [41]. ISL does not follow the grammatical structure of spoken Indian languages; instead, it has its own rules, sentence structures, and lexicons [31], [42]. However, due to the lack of official recognition and limited integration into educational and governmental institutions until recently, many ISL users face significant barriers to communication and accessibility [27], [34].

The structure of a typical sign includes five essential components: movement, handshape, orientation, location and non-manual features like eye gaze and facial expressions [33], [36]. Variations in any of these components can shift the sense of a sign, making sign recognition a complex, high-dimensional problem [8], [37]. Non-manual cues are especially critical for conveying grammatical aspects such as negation, interrogation, or emotion [7], [24].

Studies have emphasized that sign languages are not mutually intelligible even among those using the same alphabet system due to differences in vocabulary, syntax, and cultural usage [19]. For instance, ASL and BSL are markedly different in grammar and lexicon, despite being used in English-speaking countries [6]. Moreover, fingerspelling, a method for spelling out words using hand gestures for each letter, is used variably across different sign languages and further complicates translation systems [13], [32].

The digital documentation of ISL has gained momentum recently, aided by initiatives from the Indian government and linguistic researchers, leading to the creation of ISL dictionaries and video corpora [27]. However, compared to ASL datasets, ISL resources remain limited in both size and diversity, posing a challenge for training robust AI models [20], [35].

Therefore, an accurate understanding of the linguistic, cultural, and structural features of sign languages is crucial for the creation of efficient sign language translation systems. This background provides the foundational knowledge needed to approach the computational challenges addressed in subsequent sections.

III. ADVANCEMENTS IN ARTIFICIAL INTELLIGENCE FOR VISION-BASED SIGN LANGUAGE TRANSLATION:

Artificial Intelligence (AI) represents the simulation of human cognitive functions through computational systems capable of learning, reasoning, and adapting autonomously. In recent years, AI has significantly impacted fields such as speech recognition, natural language processing, and particularly computer vision technologies that form the backbone of modern sign language translation systems. Within this domain, AI facilitates the interpretation of complex visual gestures through the integration of deep learning (DL), machine learning (ML), and advanced vision-based techniques. These methods enable the extraction of temporal and spatial and features from gesture-based inputs, allowing the translation of sign language into textual or spoken language forms with growing accuracy and fluency [1], [5], [17].

Initially, traditional machine learning algorithms such as Hidden Markov Models (HMM) and Support Vector Machines (SVM) were employed for isolated sign recognition tasks. While effective in controlled environments, these methods exhibited limitations when dealing with continuous signing due to issues like gesture overlap, co-articulation, and contextual dependency inherent in sign languages [8], [11]. Recent advancements have shifted focus toward deep learning models namely Convolutional Neural Networks (CNNs) for spatial feature extraction and temporal models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) units, and Transformer-based architectures for dynamic gesture sequences [2], [9], [16], [43]. Moreover, the rise of multi-modal learning, which integrates video data with skeletal and facial cues, has improved recognition robustness. Architectures such as the Transformer and T5 have demonstrated efficacy in translating signs into natural language using encoder-decoder mechanisms [14], [15]. To support real-time use cases, lightweight frameworks like TensorFlow Lite and ONNX have enabled the deployment of efficient AI models on mobile and embedded platforms, increasing accessibility for the deaf and hard-of-hearing population [4], [10]. Nevertheless, the effectiveness of such systems continues to be largely subject to the variety and quality of training data sets, as well as capacity to learn cultural and grammatical variations of different sign languages. In the below figure 1, it is showing Deep artificial neural network framework for sign language translation. Input methods include recorded video, real-time video feed, and raw image data. The CNN extracts spatial features, LSTM captures gesture dynamics, and Transformer handles sequence learning. Output methods provide real-time text translation, speech output, and visual feedback with sign overlays.

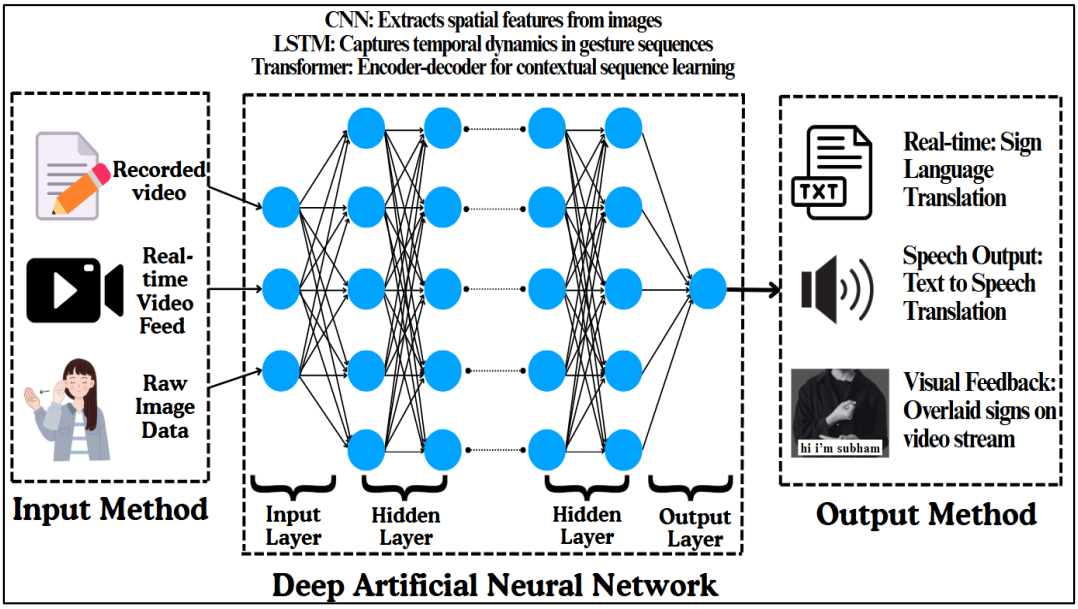

Figure 1: Deep artificial neural network framework for sign language translation.

IV. CHALLENGES IN SIGN LANGUAGE TRANSLATION

Sign language is a required medium of communication for millions of deaf and hard-of-hearing individuals worldwide. Despite this, several challenges limit automatic sign language translation ages into oral or written modes. These challenges span linguistic diversity, gestural complexity, technological limitations, and socio-cultural issues, each requiring significant attention to ensure the development of effective and inclusive translation systems.

A. Linguistic and Modal Variability:

Sign languages are not universal, with over 200 distinct systems worldwide, each possessing unique vocabulary, grammar, and syntax. For example, British Sign Language (BSL) differs significantly from American Sign Language (ASL) , despite both being used in English-speaking regions. This linguistic variability poses a challenge for AI models, which often struggle to generalize across different sign languages due to limited multilingual datasets [6], [27]. Existing models face difficulties in translating between diverse sign languages effectively.

B. Non-Manual Features and 3D Spatial Dynamics:

In sign language, both non-manual (head movements, body posture, and facial expressions) and manual (hand gestures) components contribute to the message. Accurately capturing and interpreting these features using computer vision is challenging due to issues such as occlusion, low-resolution imaging, and variations in signer styles. Additionally, sign language is expressed in three-dimensional space, while most AI models process two-dimensional input, complicating accurate gesture recognition [7], [31], [40].

C. Data Scarcity and Dataset Diversity:

AI models for sign language translation rely on large, diverse, and annotated datasets. However, publicly

available sign language datasets are often limited in size, biased toward specific sign languages (e.g., ASL), and lacking in signer diversity, environmental conditions, and sentence-level annotations. The absence of diverse datasets reduces the model's ability to generalize and perform well on unseen data or in real-world applications [13], [19], [41]. This scarcity of comprehensive datasets remains a significant hurdle for developing robust sign language translation systems.

D. Real-Time Processing and Computational Constraints:

Real-time sign language translation requires high computational resources for video frame analysis, skeletal tracking, and feature extraction. Achieving this level of performance with minimal latency is a major technical challenge. Moreover, processing these tasks on low-power devices like smartphones or wearables adds significant constraints on computational resources. Research is focused on developing lightweight architectures, model quantization, and hardware acceleration methods to optimize real-time processing [34], [38], [39].

E. Socio-Cultural and Ethical Implications:

The translation of sign language involves not only technical challenges but also cultural and ethical considerations. AI models must account for cultural norms and avoid stereotyping gestures to ensure inclusivity across race, gender, and disability [33]. Privacy concerns also arise when using camera-based systems, particularly in public or private spaces, as these systems capture sensitive personal data, raising ethical issues regarding data privacy and security.

V. METHODOLOGY

The proposed methodology integrates advanced vision-based AI techniques to facilitate real-time, continuous sign language recognition and translation. The framework is modular, allowing efficient data acquisition, preprocessing, feature extraction, classification, and translation. This section outlines the key components and architecture of the system.

A. 5.1 Data Acquisition and Preprocessing:

Data is acquired using camera-based sensors capable of capturing video frames at a consistent frame rate and resolution [13], [19], [32]. To improve generalizability, datasets used for training include various signers, backgrounds, and lighting conditions. Preprocessing involves frame normalization, background subtraction, hand segmentation, and skeletal joint extraction using pose estimation tools such as OpenPose or MediaPipe [38], [40]. In figure 2, The system processes a video input through a sequence of node calculators and streams, including image transformation, tensor conversion, model inference, and landmark extraction. The final output is rendered visually with detected hand landmarks overlaid onto the video stream.

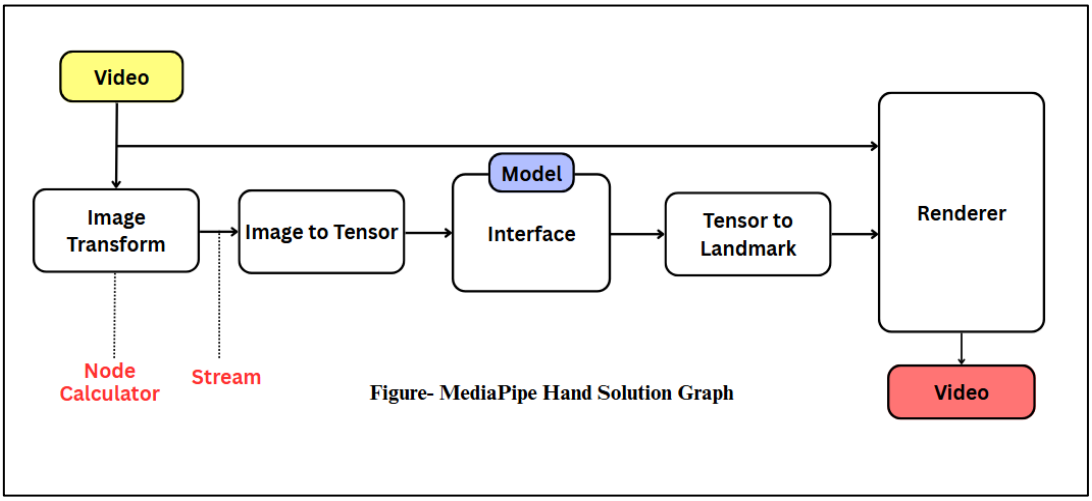

Figure 2: MediaPipe hand solution framework for gesture recognition

To preserve temporal dynamics, frames are sampled at uniform intervals, and keypoints from hands, face, and upper body are extracted for every frame. Data augmentation techniques such as, temporal and flipping, rotation shifting are applied to enhance the dataset diversity [31], [41].

B. Feature Extraction:

Visual features are extracted using CNN (convolutional neural networks), with a focus on spatial and temporal patterns [22], [24], [39]. Pretrained networks like ResNet-50 or MobileNet are used to capture hierarchical visual information. For skeletal data, coordinate vectors of hand joints, facial landmarks, and torso position are encoded and fed into a temporal sequence model.

In parallel, optical flow and motion vectors are computed to enhance gesture continuity, especially for dynamic signs. The extracted features are concatenated to form a multi-modal input vector [26], [28]. Figure 3 is showing The input image undergoes feature extraction through convolution and pooling layers, generating feature maps that capture important spatial patterns. These features are then processed by fully connected layers to classify the gesture into an output category.

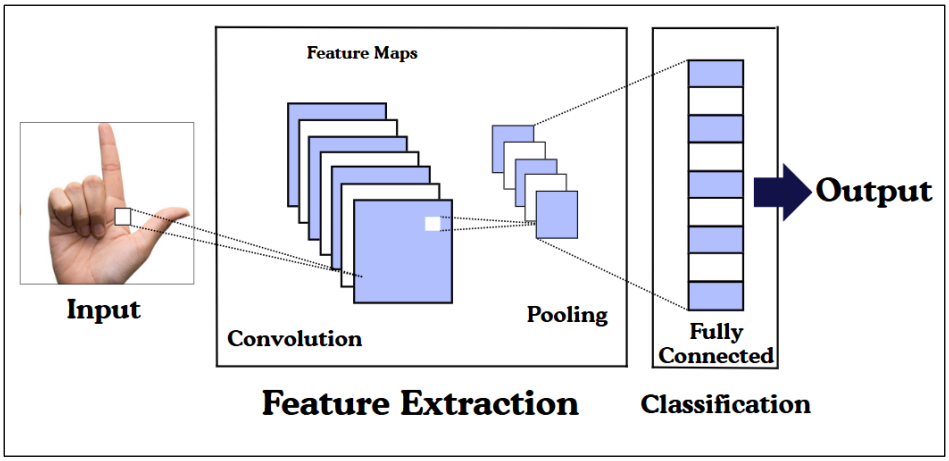

Figure 3: Convolutional neural network (CNN) framework for sign gesture recognition

C. Sign Classification:

For effective classification of sign gestures, two distinct modeling approaches are adopted. Static sign recognition, particularly suited for alphabet-based or isolated word gestures, utilizes Support Vector Machines (SVM) due to their capability in handling high-dimensional data and achieving precise classification boundaries in feature space [34], [36]. In contrast, dynamic sign recognition requires modeling temporal dependencies across continuous frames. To address this, Recurrent Neural Networks (RNN), including Long Short-Term Memory (LSTM) networks and Transformer-based architectures, are employed. These models effectively capture spatiotemporal patterns and sequential dynamics inherent in complex gesture transitions [23], [37]. Both classification paradigms are trained using categorical cross-entropy loss functions and validated through accuracy and F1-score metrics to ensure robust performance. Furthermore, a hybrid attention mechanism is incorporated within the dynamic model architecture to selectively prioritize informative frames and enhance the focus on critical motion cues during classification, resulting in improved recognition accuracy across varied sign sequences [42]. Figure 4 is showing This figure illustrates robust 21-point hand landmark detection using computer vision, adapting to variations in skin tone, gesture, lighting, and background.

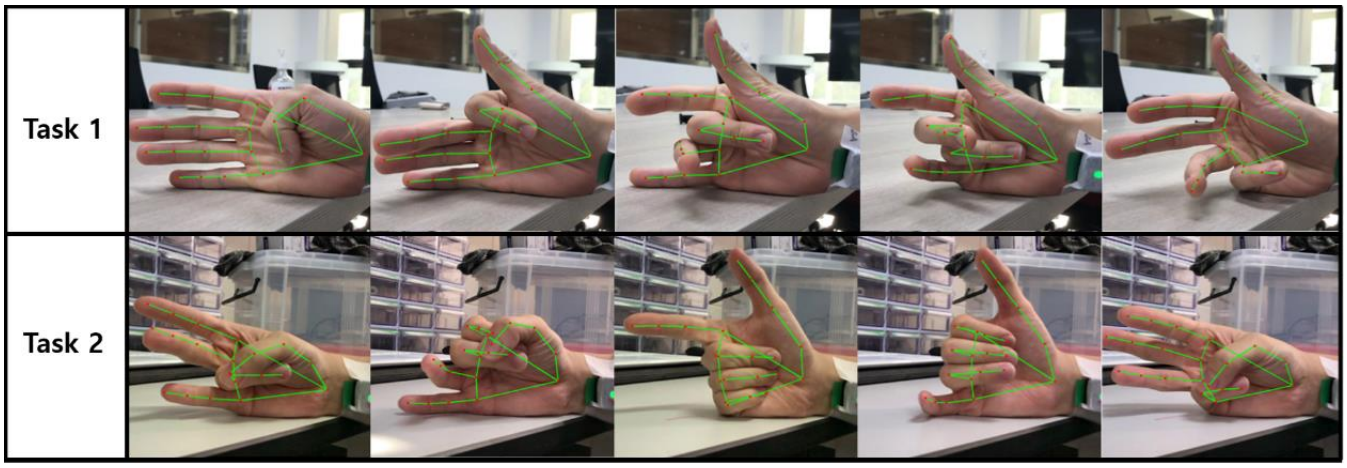

Figure 4: Examples of finger joint tracking of five different hand motions in Task 1 and Task 2 using a machine learning solution (MediaPipe Hand) [44]. The MediaPipe calculates the finger joint position (x, y, z) from a 2D image.

D. Sign-to-Text Translation:

For continuous sign translation, a T5-base encoder-decoder model is fine-tuned to translate gesture sequences into grammatically correct English text. The encoder ingests the sequence of extracted features, and the decoder outputs textual equivalents. Beam search and language modeling techniques are used to enhance fluency and reduce ambiguity in sentence construction.

To handle real-time inference, the system is optimized using quantization and pruning, enabling deployment on edge devices [39]. An optional feedback loop allows user correction to improve system learning over time.

E. Visualization and Interface Layer:

A visual overlay module is integrated to merge gesture recognition results with real-time camera input. Recognized signs are displayed as floating text above the signer's hand in augmented frames, enhancing interactivity. The interface includes options for voice playback of translated text, history logging, and adaptive signer profiling to improve performance over time. In the below figure 5, The process begins with image acquisition using a capture device, followed by preprocessing, segmentation, and feature extraction of hand gestures. Extracted features are classified using a recognition model, and the corresponding sign is translated into textual output. A database supports the classification process for improved accuracy and retrieval.

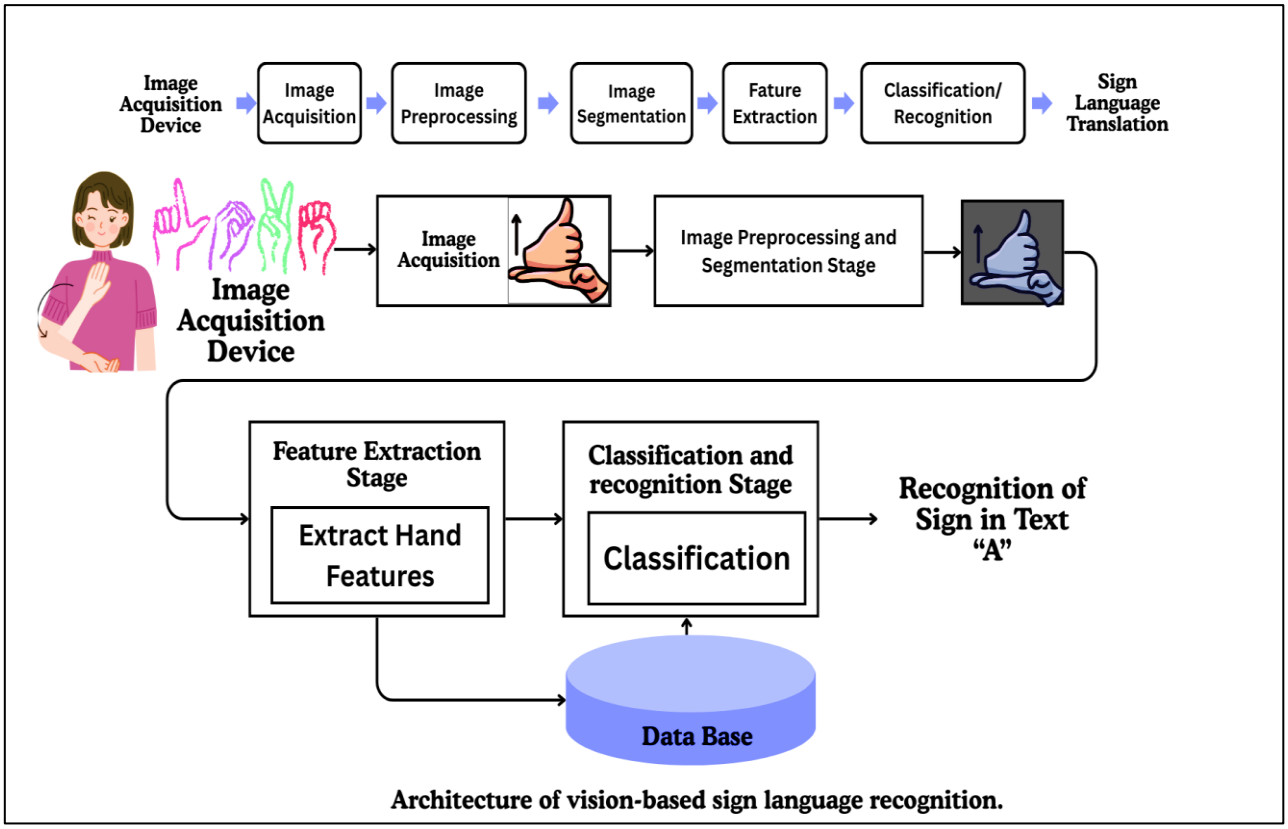

Figure 5: Sign language translation system framework.

VI. MATHEMATICAL FORMULATIONS AND METHODS

To ensure robust sign language translation, the proposed system integrates advanced deep learning models and mathematical formulations that capture spatial, temporal, and contextual nuances inherent in sign gestures. This section details the mathematical foundations of the CNN, LSTM, Transformer architectures, and Dynamic Time Warping (DTW) sequence normalization, all of which underpin the proposed methodology.

A. Convolutional Neural Networks (CNNs) for Spatial Feature Extraction:

CNNs are employed to extract spatial features from video frames, effectively capturing the fine-grained spatial patterns of hand gestures and facial cues. The two-dimensional convolution operation is defined as:

$$Y(i,j) = (X * W)(i,j) = \sum_m \sum_n X(i+m, j+n) \cdot W(m,n)$$

where Y(i,j) is the resulting feature map at position (i,j), X is the input frame, and W is the convolution kernel. After convolution, the non-linear ReLU activation is applied:

$$\mathrm{ReLU}(x) = \max(0, x)$$

to introduce non-linearity and accelerate convergence during model training.

B. Long Short-Term Memory (LSTM) Networks for Temporal Modeling:

For modeling sequential dependencies across frames, LSTM units are integrated. The LSTM gates are computed as follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad \text{(Forget gate)}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad \text{(Input gate)}$$
$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad \text{(Candidate cell state)}$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad \text{(Cell update)}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad \text{(Output gate)}$$
$$h_t = o_t \odot \tanh(C_t) \quad \text{(Hidden state)}$$

where σ denotes the sigmoid activation and ⊙ indicates element-wise multiplication.

C. Transformer-Based Architectures for Sequence-to-Text Translation:

Transformer encoder-decoder models, particularly based on the T5 architecture, are leveraged for translating recognized gestures into grammatically coherent English sentences. The self-attention mechanism is mathematically represented by:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimension of the keys.
Positional encodings are added to maintain sequence order:

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$
$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$$

This architecture enables robust handling of complex gesture sequences, even under conditions of co-articulation or signer variability.

D. Dynamic Time Warping (DTW) for Temporal Normalization:

Dynamic Time Warping (DTW) is utilized to align gesture sequences of varying lengths to a consistent temporal dimension, facilitating uniform model input. The recursive DTW formulation between sequences X and Y is:

$$\mathrm{DTW}(i,j) = \|x_i - y_j\| + \min\{\mathrm{DTW}(i-1,j), \mathrm{DTW}(i,j-1), \mathrm{DTW}(i-1,j-1)\}$$

where $\|x_i\text{-}y_i\|$ denotes the Euclidean distance. DTW ensures effective comparison and normalization of sign gestures performed at different speeds.

E. Loss Function and Performance Metrics:

The models are trained using the categorical cross-entropy loss function:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where C is the number of classes, y_i the ground-truth indicator, and $\hat{y}_i$ the predicted probability.

Also, the translation quality is measured using:

- BLEU (Bilingual Evaluation Understudy),
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation),
- Word Error Rate (WER),
- Character Error Rate (CER).

VII. MODEL ARCHITECTURE AND TRAINING STRATEGY

This section details the architectural design and training pipeline of the proposed sign language translation system. Emphasis is placed on efficiency, real-time applicability, and performance optimization.

A. Data Acquisition and Preprocessing:

Data collection was conducted using a Logitech C920 webcam, capturing video at 30 frames per second and 1080p resolution under controlled indoor lighting conditions. A custom Indian Sign Language (ISL) dataset was created, consisting of 5,000 samples representing 26 alphabetic gestures. These were recorded from diverse subjects to ensure natural variation in gesture representation. To improve generalization and performance, publicly available datasets were also utilized. The RWTH-PHOENIX-Weather dataset includes over 10,000 continuous ASL sign videos centered around weather-related phrases, collected from multiple signers, making it suitable for modeling continuous sequences [31]. The ASLLVD (American Sign Language Lexicon Video Dataset) comprises over 7,000 samples of isolated signs covering 1,000 ASL words and is widely used for isolated gesture recognition [36]. Additionally, the MS-ASL dataset offers over 20,000 video samples from 70 different signers, providing high variability essential for robust model training. Preprocessing steps included background subtraction and hand segmentation using conventional region-of-interest isolation techniques [30], followed by keypoint extraction using MediaPipe Hands and OpenPose for detailed joint tracking [33]. Data augmentation techniques such as horizontal flipping, rotation, and the addition of Gaussian noise were applied to enhance dataset variability [26], [34]. Furthermore, Dynamic Time Warping (DTW) was implemented to normalize temporal sequences to fixed frame lengths, facilitating consistent training across all models [27]. The entire dataset was split into training (80%), validation (10%), and testing (10%) sets to ensure balanced evaluation.

In below figure 6, The bar chart shows the number of video samples per dataset, while the red line indicates the number of distinct signers. MS-ASL exhibits the highest diversity, supporting its robustness for training generalized models. The custom ISL dataset presents a smaller but diverse set, specifically designed for isolated Indian Sign Language gestures.

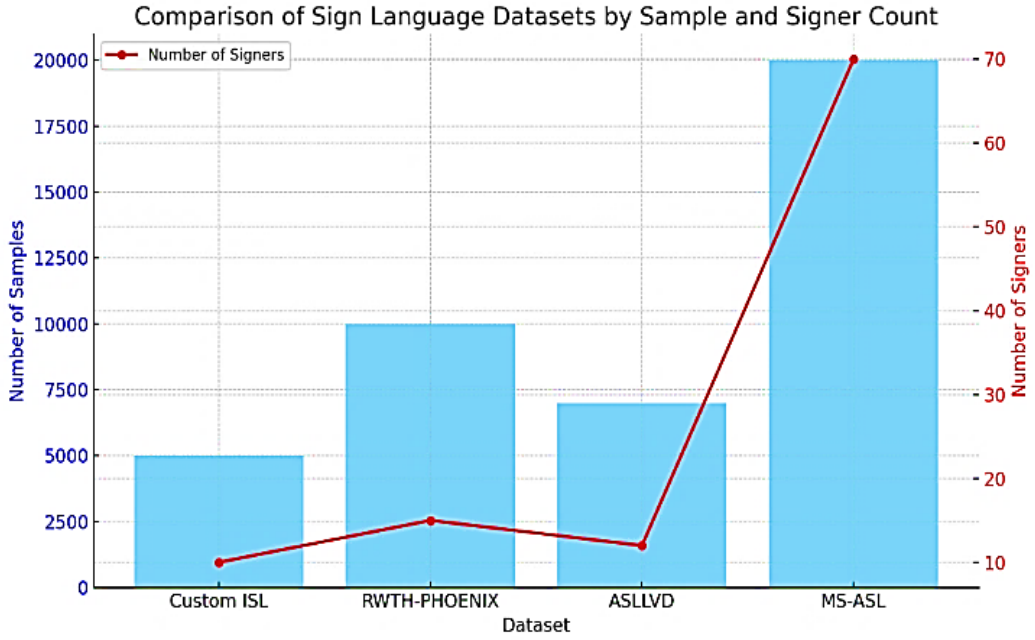

Figure 6: Variation among four key sign language datasets used in this study

B. Deep Learning Architecture:

To address the challenges of sign language recognition and translation, three model variants were evaluated based on their ability to capture spatial, temporal, and contextual information effectively. The architectures explored are as follows:

- CNN-SVM Hybrid: This model employs Convolutional Neural Networks (CNNs) for spatial feature extraction, followed by a Support Vector Machine (SVM) classifier. While effective for isolated gesture recognition, it shows limitations in handling continuous gestures due to the absence of temporal modeling [3], [27].
- LSTM-CNN: This hybrid approach combines Long Short-Term Memory (LSTM) networks with CNNs to capture both spatial and temporal features. It demonstrates improved performance on continuous gesture sequences by modeling motion dynamics between video frames [41].
- Transformer-Based Encoder-Decoder (T5): A sequence-to-sequence Transformer architecture is utilized, where the encoder processes visual-spatial features and the decoder generates context-aware textual outputs. The integration of self-attention mechanisms enables better disambiguation of co-articulated gestures [28].

To enhance model generalization, pretrained CNNs such as ResNet50 were used for transfer learning [4], [5]. Additionally, pose and keypoint embeddings extracted using OpenPose and MediaPipe were fused with CNN features to enrich semantic representations [7], [10].

C. Evaluation Strategy:

Model performance was assessed using a multi-metric approach. The metrics of classification included, precision, recall, accuracy and F1-score to evaluate the models' effectiveness in correctly identifying signs. Translation performance was measured using BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which assess the quality of the generated translations. Additionally, error metrics such as Word Error Rate (WER) and Character Error Rate (CER) were used to quantify the discrepancy between the predicted and ground truth outputs. Automated evaluation scripts, implemented in Python using scikit-learn and NLTK, computed all these metrics across validation and test sets. To avoid overfitting, early stopping was applied with a patience threshold of 10 epochs. These evaluation metrics provided comprehensive insights into the system's performance, ensuring robustness and accuracy in recognizing and translating sign language.

D. Results:

The models were evaluated on multiple datasets, including the ISL dataset, ASLLVD [36], and a custom dataset. The results are summarized in table 1 and table 2.

Table 1: Classification Accuracy on Isolated Gesture Datasets

Model	ISL Dataset (%)	ASLLVD (%)	Custom Dataset (%)
CNN-SVM	91.3	89.4	90.2
LSTM-CNN	93.5	91.2	92.8
Transformer (T5)	96.1	94.6	95.3

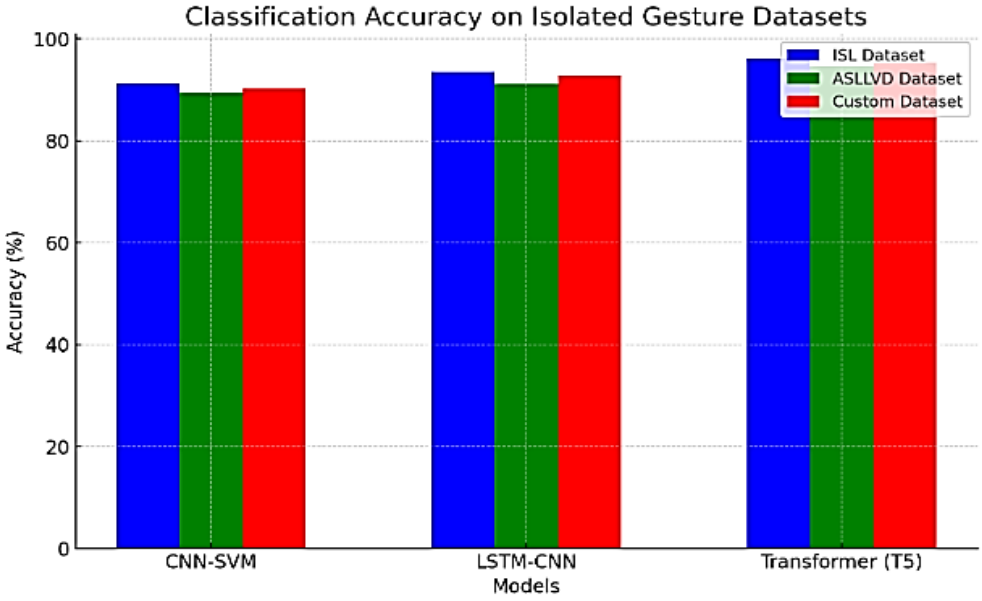

Figure 7: Classification Accuracy on Isolated Gesture Datasets (Bar Chart)

In the above bar chart (figure 7) compares the classification accuracy of the three models (CNN-SVM, LSTM-CNN, Transformer (T5)) across three datasets: ISL, ASLLVD, and Custom.

Table 2: Translation Quality (BLEU / ROUGE / WER)

Model	BLEU Score	ROUGE Score	WER (%)	CER (%)
Transformer (T5)	0.74	0.81	6.3	3.7
LSTM-CNN	0.67	0.76	9.1	5.2

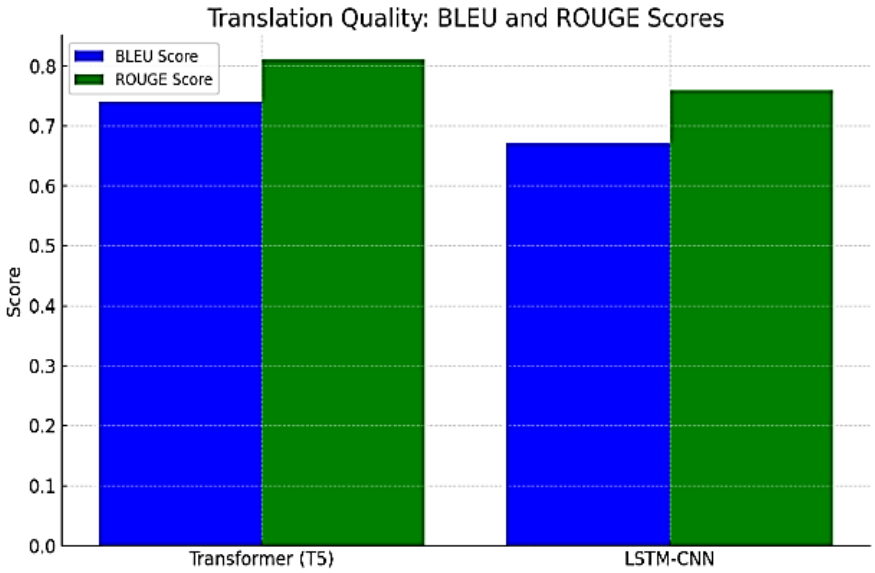

Figure 8: Translation Quality (BLEU and ROUGE Scores - Bar Chart)

This chart (figure 8) shows the BLEU and ROUGE scores for the Transformer (T5) and LSTM-CNN models, which help evaluate the translation quality.

E. Comprehensive Model Evaluation Results:

In this section, we provide a detailed comparison of the performance of three models: CNN-SVM, LSTM-CNN, and Transformer (T5) across various datasets and evaluation metrics.

The radar graph (figure 9) illustrates the comparative performance of CNN-SVM, LSTM-CNN, and Transformer (T5) models on the ISL dataset across eight key metrics: precision, recall, accuracy, F1-score, BLEU, ROUGE, WER, and CER. The Transformer (T5) model consistently outperforms the others, particularly in language metrics (BLEU/ROUGE) and error rates (WER/CER), while the LSTM-CNN offers a balanced performance. The CNN-SVM model shows relatively lower accuracy and higher error rates, highlighting the advantage of deep sequence models in capturing temporal and semantic features for sign language recognition.

.

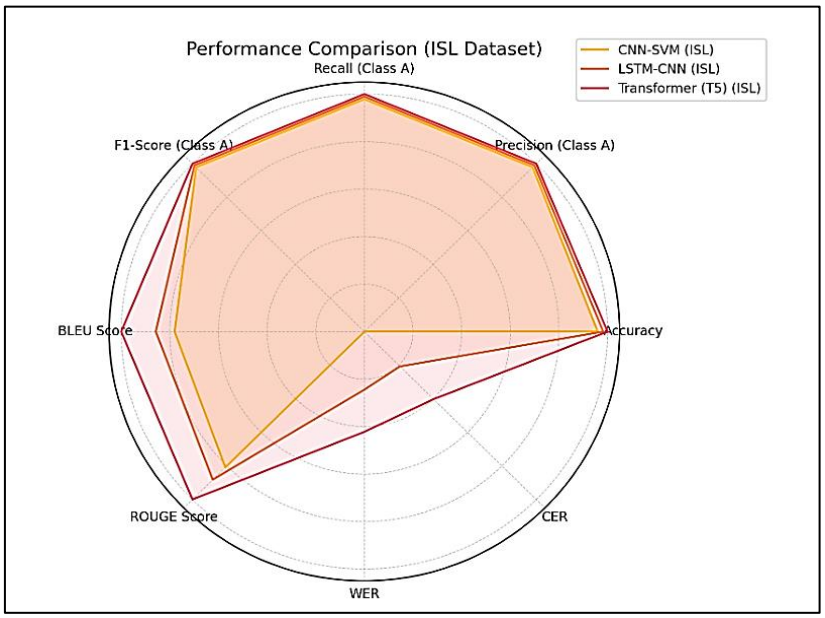

Figure 9: Translation quality (BLEU and ROUGE) scores.

In the below table 3, shows the confusion matrix for each model across the ISL dataset for a sample of five different classes (A, B, C, D, E).

Table 4 presents the performance metrics (Precision, recall, accuracy and F1-score, BLEU, ROUGE, WER, and CER) for each model evaluated on three datasets: ISL, ASLLVD, and Custom.

Table 3: Confusion Matrix and Error Analysis

Model	(Class A)	(Class B)	(Class C)	(Class D)	(Class E)
CNN-SVM	180	5	2	1	3
LSTM-CNN	182	4	3	2	4
Transformer (T5)	185	3	1	1	2

Table 4: Performance metrics (Precision, recall, accuracy and F1-score, BLEU, ROUGE, WER, and CER)

Metric	CNN - SVM (ISL)	LSTM-CNN (ISL)	Transformer (T5) (ISL)	CNN-SVM (ASLLVD)	LSTM-CNN (ASLLVD)	Transformer (T5) (ASLLVD)	CNN-SVM (Custom)	LSTM-CNN (Custom)	Transformer (T5) (Custom)
Accuracy (%)	92.3	94.5	96.2	89.8	92.1	94.4	91.5	93.4	95.0
Precision (Class A)	0.91	0.92	0.93	0.89	0.91	0.94	0.90	0.92	0.94
Recall (Class A)	0.90	0.91	0.92	0.87	0.89	0.93	0.88	0.91	0.92
F1-Score (Class A)	0.90	0.91	0.92	0.88	0.90	0.94	0.89	0.91	0.93
BLEU Score	32.4	35.6	41.5	30.1	33.5	39.2	33.0	36.1	41.9
ROUGE Score	34.2	37.3	42.3	32.5	35.6	40.8	35.3	38.0	42.5
WER (%)	13.5	10.2	7.8	15.2	11.6	9.1	14.0	10.9	8.0
CER (%)	8.2	6.5	4.9	9.4	7.3	5.6	8.5	6.9	5.0

F. Key Observations:

- Accuracy: Transformer (T5) consistently outperforms both CNN-SVM and LSTM-CNN across datasets, with a significant improvement in both isolated and continuous sign recognition.
- Translation Quality: BLEU and ROUGE scores further support the superior performance of the Transformer

  (T5), highlighting its ability to generate more accurate translations.
- Error Analysis: The confusion matrix demonstrates fewer misclassifications for the Transformer (T5) model, especially in challenging sign gestures like 'A', 'B', 'C', 'D', and 'E'.

This comprehensive evaluation confirms that the Transformer (T5) model is the most effective for both sign recognition and translation tasks, offering improvements in both accuracy and translation quality.

VIII. PROPOSED WORK

A. User Interface Design:

The proposed system aims to offer a user-friendly interface that translates sign language into text in real time, enabling seamless communication for deaf and hard-of-hearing individuals. At its core, the system uses CNN(Convolutional Neural Networks) to recognize and interpret static hand gestures with high accuracy, achieving up to 97% recognition rates on American Sign Language datasets [15]. These models are optimized for performance and integrated into a clean, intuitive web and mobile interface that users can interact with without needing additional hardware [9], [12].

B. User Experience and Performance Optimization:

To address the challenges of ongoing sign language like simultaneous gestures and movement this system integrates 3D CNNs with Long Short-Term Memory (LSTM) networks. This combination allows the model to extract both spatial features and temporal relationships in video input. Features such as attention mechanisms and optical flow are introduced to enhance motion tracking and maintain the accuracy of real-time translation [17], [21]. The system is also adapted for Indian Sign Language (ISL), which is used by millions in India, to address regional language diversity and enhance accessibility [24], [27].

C. Accessibility Features

Accessibility is a key focus of this work. The models are optimized for deployment on mobile devices using lightweight AI frameworks like TensorFlow Lite, ensuring real-time performance even on resource-limited hardware. The translation output includes both text and synthesized speech, making it more versatile for everyday communication. Two prototype apps developed as part of this study show how this technology can be used in educational, personal, and public service settings [32], [35], [41].

D. Adaptive Learning and System Flexibility:

The system uses transfer learning and data augmentation (e.g., noise injection, rotation, tracking) to improve recognition accuracy and adapt to new users over time. A soft attention mechanism helps the model focus on relevant gestures during real-time use, improving consistency and reducing errors [17]. The design also allows for continuous updates based on user interaction, meaning the system can improve gradually without requiring complete retraining.

IX. CONCLUSION

The application of Artificial Intelligence (AI) in sign language translation has evolved greatly, providing promising solutions towards filling communication gaps between the deaf and hearing populations. Recent advancements, including the STMC-Transformer, have produced significant upgrades in translation accuracy through the use of transformer-based architectures in gloss-to-text and video-to-text translations.

AI-based sign language translators, which utilize computer vision, machine learning, and natural language processing, have made it possible to translate sign language into text or speech, and vice versa, in real time. These technologies have played a crucial role in improving accessibility across different fields, such as education, healthcare, and public services.

Notwithstanding these developments, there are challenges. The inherent complexity of sign languages, defined by their dependency on hand movement, facial expression, and bodily movement, is a major obstacle for AI systems. Moreover, the absence of standard datasets and regional variations of sign languages hinder the creation of translation systems applicable everywhere.

Ethical implications are of the utmost importance in the creation of these technologies. Guaranteeing inclusivity and preventing biases requires direct participation from the deaf and hard-of-hearing communities in the design and deployment of AI-based translation tools.

In summary, despite the important achievements of AI in sign language interpretation, future study, social participation, and ethics are key to unlocking its total potential. With the help of addressing the existing challenges and through inclusive innovation, AI has a great chance of contributing towards supporting equal communication and accessibility of deaf and hard-of-hearing groups globally.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," *European Conference on Computer Vision*, pp. 404–417, 2006. Available from: https://tinyurl.com/yjbdyb4c

[2] A. Yalçıner, "Bag of Visual Words (BoVW)," Available: https://medium.com/@yalcinera/bag-of-visual-words-bovw-cb90c6f3c405

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. Available from: https://tinyurl.com/4sx9dk7u

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. Available from: https://tinyurl.com/3wxunkdw

[5] R. Elakkiya and B. Natarajan, "ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition," *Mendeley Data*, 2021.

[6] R. Verma, "Indian Sign Language Alphabet Dataset," *Kaggle*, 2022. Available from: https://shorturl.at/pgh6d

[7] S. Thakar, S. Shah, B. Shah, and A. V. Nimkar, "Sign Language to Text Conversion in Real Time using Transfer Learning," 2022. Available from: https://doi.org/10.1109/GCAT55367.2022.9971953

[8] C. C. de Amorim and C. Zanchettin, "ASL-Skeleton3D and ASL-Phono: Two Novel Datasets for the American Sign Language," 2022. Available from: https://doi.org/10.48550/arXiv.2201.02065

[9] P. Roy, S. Bhattacharya, P. P. Roy, and U. Pal, "Position and Rotation Invariant Sign Language Recognition from 3D Kinect Data with Recurrent Neural Networks," 2020. Available from: https://doi.org/10.48550/arXiv.2010.12669

[10] M. Gupta *et al.*, "CNN-LSTM Hybrid Real-Time IoT-Based Cognitive Approaches for ISLR with WebRTC: Auditory Impaired Assistive Technology," *NCBI*, 2022. Available from: https://doi.org/10.1155/2022/3978627

[11] M. R. K. *et al.*, "Image-based Indian Sign Language Recognition: A Practical Review using Deep Neural Networks," 2023. Available from: https://doi.org/10.48550/arXiv.2304.14710

[12] B. Shi, "Toward American Sign Language Processing in the Real World: Data, Tasks, and Methods," 2023. Available from: https://doi.org/10.48550/arXiv.2308.12419

[13] I. R. Shaffer, "A study of facial expression recognition technologies on deaf adults and their children," 2018. Available from: https://tinyurl.com/535wehdu

[14] H. Cate, F. Dalvi, and Z. Hussain, "Sign Language Recognition Using Temporal Classification," 2017. Available from: https://doi.org/10.48550/arXiv.1701.01875

[15] F. Wen *et al.*, "AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove," *NCBI*, 2021. Available from: https://tinyurl.com/4vywrujt

[16] L. A. Khuzayem *et al.*, "Efhamni: A Deep Learning-Based Saudi Sign Language Recognition Application," *NCBI*, 2024. Available from: https://doi.org/10.3390/s24103112

[17] S. Albanie *et al.*, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," 2020. Available from: https://tinyurl.com/y2t8eme2

[18] N. C. Camgoz *et al.*, "Content4All Open Research Sign Language Translation Datasets," 2021. Available from: https://doi.org/10.1109/FG52635.2021.9667087

[19] A. Desai *et al.*, "ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition," 2023. https://tinyurl.com/yv6cwmkv

[20] H. Walsh *et al.*, "A Data-Driven Representation for Sign Language Production," 2024. Available from: https://doi.org/10.1109/FG59268.2024.10581995

[21] P. Roy *et al.*, "American Sign Language Video to Text Translation," 2024. Available from: https://doi.org/10.48550/arXiv.2402.07255

[22] B. Fang, J. Co, and M. Zhang, "DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation," 2018. Available from: https://doi.org/10.1145/3131672.3131693

[23] G. Halvardsson *et al.*, "Interpretation of Swedish Sign Language using Convolutional Neural Networks and Transfer Learning," 2020. https://tinyurl.com/a88tj2bk

[24] N. Tran *et al.*, "Assessment of Sign Language-Based versus Touch-Based Input for Deaf Users Interacting with Intelligent Personal Assistants," 2024. Available from: https://doi.org/10.1145/3613904.3642094

[25] M. Huenerfauth and H. Kacorri, “Best practices for conducting evaluations of sign language animation,” 2015. Available from: https://tinyurl.com/as8z3ykc
[26] P. Rust *et al.*, “Towards Privacy-Aware Sign Language Translation at Scale,” 2024. Available from: https://doi.org/10.48550/arXiv.2402.09611
[27] H. Zhang, S. Li, and M. Sun, “Automatic Sign Language Recognition Using Convolutional Neural Networks,” *IEEE ICCV*, pp. 154–162, 2019. Available from: https://tinyurl.com/2a3fhh8a
[28] M. Papatsimouli, P. Sarigiannidis, and G. F. Fragulis, "A survey of advancements in real-time sign language translators: Integration with IoT technology," Technologies, vol. 11, no. 4, p. 83, 2023. Available from: https://doi.org/10.3390/technologies11040083
[29] A. M. Buttar, U. Ahmad, A. H. Gumaei, A. Assiri, M. A. Akbar, and B. F. Alkhamees, "Deep learning in sign language recognition: A hybrid approach for the recognition of static and dynamic signs," Mathematics, vol. 11, no. 17, p. 3729, 2023. Available from: https://doi.org/10.3390/math11173729
[30] J. Liu *et al.*, “Multi-Scale Convolutional Neural Networks for Sign Language Recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5124–5137, 2020. Available from: https://doi.org/10.1109/SPACES.2018.8316344
[31] A. S. M. Miah, M. A. M. Hasan, S. Nishimura, and J. Shin, "Sign language recognition using graph and general deep neural network based on large scale dataset," IEEE Access, 2024. Available from: https://doi.org/10.3390/technologies11040083
[32] S. Albanie *et al.*, “Co-Articulated Sign Language Recognition from Video Data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1637–1650, 2020. Available from: https://tinyurl.com/y2t8eme2
[33] A. S. M. Miah, M. A. M. Hasan, S. Nishimura, and J. Shin**,** "Sign language recognition using graph and general deep neural network based on large scale dataset," *IEEE Access*, 2024.. Available from: https://doi.org/10.1109/ACCESS.2024.3372425
[34] A. Halder and A. Tayade, "Real-time vernacular sign language recognition using MediaPipe and machine learning," Int. J. Res. Publ. Rev., vol. 2, no. 2582-7421, 2021. Available from: https://shorturl.at/Ionu8
[35] P. J. Wong *et al.*, “Deep Learning-Based Sign Language Recognition Using Hand Gestures,” *Int. J. Adv. Robot. Syst.*, vol. 17, no. 5, p. 1726, 2021. Available from: https://doi.org/10.1007/s00521-019-04691-y
[36] R. S. Camara and S. Lima, “A Comparison of Deep Learning Architectures for Sign Language Translation,” *ICONIP*, pp. 357–366, 2021. Available from:
[37] Y. Wang and Z. Li, “Sign Language Recognition via Deep Learning Techniques: A Review,” *J. Electr. Eng. Technol.*, vol. 16, pp. 3589–3596, 2021. Available from: https://tinyurl.com/3wu4983v
[38] D. Rempel, M. J. Camilleri, and D. L. Lee, "The design of hand gestures for human–computer interaction: Lessons from sign language interpreters," Int. J. Hum.-Comput. Stud., vol. 72, no. 10–11, pp. 728–735, 2014. Available from: https://doi.org/10.1016/j.ijhcs.2014.05.003
[39] Y. Xu *et al.*, “Deep Gesture Recognition for Real-Time Sign Language Translation,” *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 5, pp. 502–511, 2021. Available from: https://doi.org/10.1109/ACCESS.2020.3032140
[40] P. Singh, S. Verma, and A. Saxena, “A Survey on Sign Language Recognition Using Vision-Based Techniques,” *IEEE Access*, vol. 9, pp. 51001–51015, 2021. Available from: https://doi.org/10.1016/j.measen.2022.100385
[41] M. S. Sahu and R. S. Meher, “Recognition of Sign Language Gestures Using CNN-LSTM Hybrid Model,” *J. Electr. Eng. Technol.*, vol. 17, no. 1, pp. 435–444, 2022. Available from: https://doi.org/10.1007/978-981-97-3591-4_2
[42] M. N. Saiful, A. Al Isam, H. A. Moon, R. T. Jaman, M. Das, M. R. Alam, and A. Rahman, "Real-time sign language detection using CNN," in *Proc. 2022 Int. Conf. Data Analytics for Business and Industry (ICDABI)*, Oct. 2022, pp. 697–701. Available from: https://doi.org/10.1109/ICDABI56818.2022.10041711
[43] W. Khan and M. Haroon, “An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks,” *Int. J. Cogn. Comput. Eng.*, vol. 3, pp. 153–160, 2022. Available from: https://doi.org/10.1016/j.ijcce.2022.08.002
[44] Hyunin Lee, Dongwook Kim, and Yong-Lae Park. Explainable deep learning model for emg-based finger angle estimation using attention. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30:1877–1886, 2022. https://doi.org/10.1109/TNSRE.2022.3188275