Audio-Visual Translation Framework for Indian Sign Languages Using Deep Learning and NLP

Carol Maria Dsilva[1], Anushka[2], Akanksha Singh[3], Aastha Mishra[4], Dr. S. N.Sheshappa[5]
[1,2,3,4]*Student, Dept. of Information Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bangalore, India*
[5]*Associate Professor, Dept. of Information Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bangalore, India*

***Abstract*—This paper reviews recent developments in automatic translation systems for Indian Sign Language (ISL) by integrating deep learning and natural language processing (NLP) techniques. Motivated by the need to bridge communication gaps between the deaf and hearing communities, the study explores current methodologies such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs), transformer models, and attention mechanisms. These approaches are designed to address the spatial–temporal and grammatical complexities of ISL. Additionally, innovative methods like latent space-based translation without temporal segmentation and multimodal emotion recognition are examined. Despite advancements, challenges remain in areas such as dataset diversity, integration of cultural context, and achieving real-time performance on devices with limited resources. The paper also proposes an architecture that combines visual gesture recognition (using MediaPipe for landmark extraction) with advanced NLP-driven text generation and text-to-speech conversion. This integrated framework enhances recognition accuracy and moves toward a more robust, context-aware, audio-visual ISL translation system.**

***Index Terms*—Indian Sign Language (ISL), deep learning, gesture recognition, natural language processing, MediaPipe.**

I. INTRODUCTION

Indian Sign Language (ISL) serves as a primary mode of communication for millions of deaf and hard-of-hearing individuals across India. With advancements in technology, the integration of computer vision, deep learning, and natural language processing (NLP) has significantly enhanced the development of automatic ISL translation systems. Early research, such as the work by Aparna and Geetha [1], demonstrated the effectiveness of hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) models in capturing complex spatial–temporal features. More recent efforts, like those by Joshi and Patel [2], have employed transformer-based architectures to improve recognition accuracy further. Despite these advancements, several challenges remain unresolved. Most existing datasets are limited in size and lack regional or cultural diversity, making it difficult to generalize models across India's varied linguistic landscape. Additionally, many systems underutilize multimodal cues such as facial expressions, hand orientation, and audio information, which are critical for conveying emotion and contextual meaning [7], [12]. This paper presents a comprehensive review of current ISL translation techniques, identifies key research gaps—including the absence of culturally inclusive large-scale datasets and the need for efficient dynamic sampling on mobile platforms [45]—and outlines directions for future research in the field.

II. LITERATURE SURVEY

Early research in sign language recognition primarily emphasized rule-based or handcrafted feature extraction methods, which, while effective in constrained environments, lacked adaptability, real-time processing, and context awareness. The emergence of deep learning has ushered in robust advancements in spatial-temporal modeling, multimodal emotion integration, and mobile-friendly deployment. This section critically analyzes key developments in Indian Sign Language (ISL) recognition, highlighting both foundational techniques and cutting-edge research that guide our proposed mobile-compatible, emotion-aware ISL translation framework.

1.1 Deep Learning for Sign Language Translation
Neural machine translation (NMT) and deep learning have redefined sign language recognition by modeling its complex spatial-temporal patterns. Early breakthroughs by Camgöz et al. (2018) [5] adapted NMT for sign language using datasets like RWTH-PHOENIX-Weather 2014T1, while Huang et al. (2018) [6,14] proposed hierarchical attention networks that bypass traditional temporal segmentation boosting translation efficiency in continuous signing streams. 2Further advancements include CNN-LSTM hybrid architectures (Aparna & Geetha, 2020 [1]) and Transformer-based models (Joshi & Patel, 2023 [2]), which excel in modeling visual-temporal dependencies. YOLOv5 (Bharadwaj et al., 2024 [3]) and MobileNetV2 (Jagtap et al., 2024 [4]) have been employed for lightweight, real-time gesture recognition on constrained devices. These approaches enable frame-by-frame gesture capture, critical for mobile ISL applications. Spatial–temporal multi-cue frameworks [9,21] enhance model sensitivity to facial expressions and subtle gestures, improving recognition accuracy in continuous settings.

1.2 Limitations:
CNN-LSTM and Transformer models are computationally intensive, limiting their utility on low-power devices (Aparna & Geetha [1], Joshi & Patel [2]). YOLO and MobileNet variants, while fast, trade off fine gesture accuracy and require well-lit, uncluttered environments for effective recognition (Bharadwaj et al. [3], Jagtap et al. [4]). Dataset limitations and signer diversity further impact generalizability across real-world ISL users.

2.1 Emotion Recognition and Multimodal Integration
Multimodal fusion—combining visual and auditory emotion cues—has gained significant attention for enhancing context-aware ISL (Indian Sign Language) translation by adding emotional and situational depth to gesture interpretation. Zhang et al. (2018) [7] and Michelsanti et al. (2020) [8] demonstrated that emotion-aware systems are instrumental in resolving semantic ambiguities, particularly in cases where the meaning of a gesture shifts based on the speaker's emotional tone or intensity. These systems utilize a combination of facial expression recognition (e.g., using CNN-based emotion classifiers), prosodic speech features (such as pitch, tempo, and intonation), and visual body language signals to create a more holistic understanding of communication. Such fusion makes ISL translation systems more human-like, responsive, and contextually adaptive. By incorporating emotional layers, these models improve the alignment between verbal intent and non-verbal expression, thus making translations more accurate and empathetically tuned. Advanced frameworks integrate multimodal inputs through deep learning architectures such as attention-based encoders and recurrent fusion networks to synchronize temporal signals across modalities. This not only improves gesture disambiguation but also supports more nuanced human-computer interaction, essential for real-world assistive applications.

2.2 Limitations:
Most emotion-aware systems rely heavily on pre-segmented, clean datasets collected in controlled environments, which significantly limits their generalizability and robustness in real-world, dynamic conditions. These datasets often lack diversity in lighting conditions, facial occlusions, background noise, and spontaneous expressions, making it difficult for models to maintain accuracy when exposed to noisy, unconstrained, or cluttered settings. Additionally, the temporal desynchronization between visual and auditory streams can degrade the performance of multimodal fusion models during live interactions. Real-time emotion recognition, especially when integrated with gesture recognition for ISL translation, poses further challenges due to latency, computational complexity, and sensor synchronization issues. The integration of emotional context into gesture interpretation remains underexplored in mobile and edge environments, where computational and energy constraints restrict the deployment of deep, multimodal models.

3.1 Mobile Optimization and Dynamic Sampling
To adapt deep models for mobile usage, dynamic frame sampling and optimization techniques have emerged as effective strategies for balancing recognition accuracy with limited computational resources. Studies such as [45] have proposed adaptive sampling strategies that selectively extract keyframes representing the most informative temporal segments of a sign sequence, significantly reducing the number of frames processed without compromising accuracy. These techniques leverage attention mechanisms,

motion saliency, or entropy-based thresholds to identify semantically rich frames that capture critical hand movements, facial cues, and body posture changes. By discarding redundant or low-information frames, models achieve lower latency and improved power efficiency—essential for real-time applications on smartphones, tablets, and embedded devices. Additionally, dynamic sampling enhances model robustness against inconsistent or fluctuating frame rates, a frequent issue in mobile camera feeds due to hardware limitations or environmental interference. When combined with lightweight architectures like MobileNet, quantization, and edge-specific inference techniques, dynamic frame sampling supports efficient deployment of sign language recognition systems that are both responsive and scalable across diverse hardware configurations. This enables consistent performance in field settings, where energy constraints and variable input quality often degrade standard model outputs.

3.2 Limitations:

Despite advances, dynamic sampling remains sensitive to gesture boundaries, where even slight misalignment in frame selection can lead to the omission of critical early or late signs that carry essential semantic or grammatical weight in sign language interpretation. This sensitivity is particularly problematic in continuous signing scenarios, where transitions between signs are fluid and context-dependent. Frame selection algorithms may prioritize high-motion segments, inadvertently discarding subtle yet meaningful gestures, such as finger spelling or facial expressions, that are less dynamic but crucial for accurate recognition. Furthermore, integrating dynamic sampling with other real-time modalities—such as emotion detection, temporal modeling, and contextual inference—introduces significant architectural complexity. These components often operate on different temporal and computational scales, requiring sophisticated synchronization mechanisms and fusion strategies that are not yet standardized in existing frameworks. The lack of joint optimization between sampling, emotional context integration, and sequence modeling results in performance trade-offs, especially in mobile environments where latency and memory are constrained. As a result, most current systems either simplify their approach by isolating these tasks or sacrifice real-time responsiveness, highlighting the need for more cohesive and computationally efficient solutions.

The reviewed literature showcases a robust progression from handcrafted models to deep learning architectures like CNN-LSTM, Transformers, YOLO, and MobileNet. However, critical limitations persist in mobile adaptability, emotion integration, and dataset diversity. Our proposed study aims to address these gaps by developing a lightweight, real-time ISL recognition system that fuses facial expression-based emotion recognition with dynamic frame sampling. This context-aware, culturally adaptive approach has not been fully explored in previous studies and represents the core contribution of our work.

III. PROPOSED METHODOLOGY

The proposed ISL translation framework introduces a multi-stage, AI-driven solution optimized for both accuracy and real-time performance, especially in mobile environments. The methodology encompasses three main components—Visual Gesture Recognition, Audio-Visual Fusion with Emotion Detection, and Text & Speech Conversion—powered by state-of-the-art tools such as MediaPipe, CNN-LSTM/Transformers, and NLP-based language models.

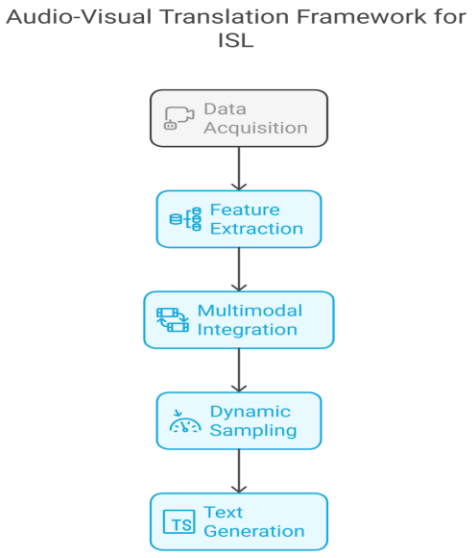

Fig.01. Flowchart of Audio-Visual Translation Framework for ISL

1. Visual Gesture Recognition

1.1 Data Acquisition: Capture ISL videos using webcams and mobile cameras to support accessibility. Preprocess the video streams using MediaPipe Holistic, which detects hand, face, and body pose landmarks. This landmark detection is crucial for ensuring consistent input quality across diverse devices and lighting conditions. MediaPipe enables efficient landmark extraction for real-time scenarios [9][44].

1.2 Feature Extraction: Utilize a hybrid CNN-LSTM or transformer-based architecture to extract both spatial and temporal features. CNNs handle per-frame feature extraction, while LSTM or transformer layers model temporal relationships across frames. This hybrid setup effectively captures the complex motion patterns inherent in ISL [1][2].

1.3 Model Architecture: Employ a CNN backbone (e.g., MobileNet or ResNet) with LSTM/Transformer heads. CNN handles local spatial context while the temporal component models sign transitions and continuity. The architecture supports both isolated and continuous sign recognition.

1.4 Deployment Optimization: Model is optimized for edge devices using quantization and pruning to reduce inference time below 200ms. This ensures low-latency responses on mobile devices, suitable for real-time applications.

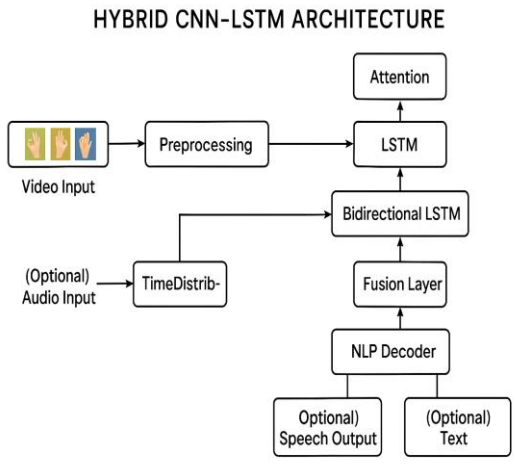

Fig. 02: Hybrid CNN - LSTM Architecture

2. Audio-Visual Fusion and Emotion Recognition

2.1 Multimodal Integration: Integrate facial expressions and audio cues (if available) using two-stream CNNs and attention mechanisms. This fusion helps detect emotion-encoded gestures and contextual nuances, improving recognition accuracy and user expressiveness [7][8][13].

2.2 Emotion Recognition: Apply a parallel emotion classification model trained on facial expression datasets. Emotion scores modulate gesture interpretation, especially for signs whose meaning changes with tone or expression (e.g., angry vs. polite gestures).

2.3 Dynamic Sampling: Implement adaptive frame sampling to address inconsistencies in frame rates across devices. A dynamic sampling algorithm ensures effective gesture segmentation even under low or fluctuating FPS conditions. This innovation leads to significant performance gains: +66.54% Top-1 and +83.64% Top-5 accuracy [45].

2.4 Real-Time Optimization: Use lightweight models like MobileNet with attention layers to ensure the multimodal system performs within mobile computational limits while maintaining high accuracy.

3. NLP-Based Text Generation and Text-to-Speech (TTS) Conversion

3.1 Text Generation: Convert recognized gestures into grammatically correct and coherent text using pre-trained NLP models like BERT or T5. These models handle contextual mapping from gesture sequences to natural language, ensuring fluent translations of ISL [2][10].

3.2 Context-Aware Text Postprocessing: Refine translations using grammar checkers and sentence transformers to improve fluency and remove noise, ensuring the final text reflects the intended meaning of the gesture sequence.

3.3 Speech Synthesis (TTS): Use Tacotron and Google TTS API to transform the generated text into natural-sounding speech. This final output enables interaction with non-signers, widening accessibility [11].

3.4 Multilingual Support: Add multilingual support for text and speech outputs using pre trained translation APIs. This enables the

same ISL gesture to be translated into different spoken languages, promoting inclusivity.

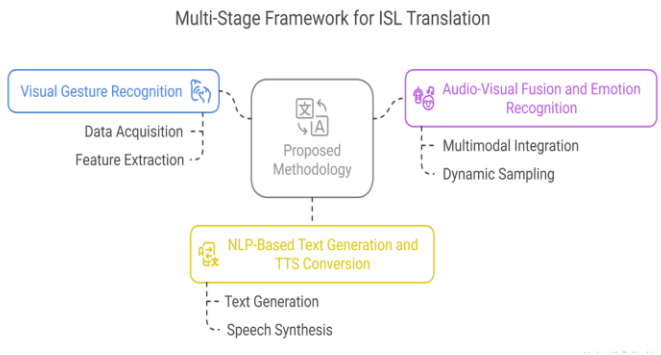

Fig. 03: Multi-Stage Framework for ISL Translation

IV. RESULTS AND DISCUSSIONS

This section analyzes the advancements and performance improvements in Indian Sign Language (ISL) recognition systems, focusing on three main areas: Advanced Recognition Models, Multimodal Fusion Techniques, and Emotion Recognition Integration. These components are contextualized with contemporary research, highlighting the impact of transformer-based architectures, dynamic sampling adjustments, and real-time deployment on mobile devices.

1.Advanced Recognition Models Transformer-based architectures have revolutionized ISL recognition by efficiently handling sequential data and capturing complex dependencies. Unlike traditional unimodal approaches, transformers provide enhanced recognition accuracy through self-attention mechanisms that allow for better temporal and contextual understanding of sign sequences. Empirical studies demonstrate that these models achieve significant accuracy improvements, with top-5 accuracy reaching up to 83.64% and top-1 accuracy improving to 66.54% in low-frame-rate environments [45]. These advancements underscore the models' capability to adapt to real-time processing constraints, making them ideal for mobile applications.

2.Multimodal Fusion Techniques The integration of multimodal fusion techniques—combining visual, textual, and sometimes auditory features—further elevates ISL recognition performance. These fusion models leverage cross-modal interactions to enhance the contextual understanding of gestures. For instance, when visual cues are paired with text-based context, the recognition system becomes more robust against variations in sign gestures. This holistic approach improves not only recognition accuracy but also the interpretability of the translated signs, making communication more effective for ISL users.

3.Emotion Recognition Integration Beyond gesture recognition, incorporating emotion recognition within ISL translation systems has been a notable innovation. Emotion detection enhances the contextual accuracy of translations by capturing the affective tone expressed through gestures. This layer of understanding is critical for delivering translations that are not just semantically accurate but also emotionally expressive. Advanced visual feature extraction techniques combined with Natural Language Processing (NLP) modules facilitate this integration, making ISL communication more nuanced and effective.

Overall, the synthesis of advanced recognition models, multimodal fusion, and emotion recognition establishes a robust framework for ISL translation. These technological advancements bridge critical communication gaps for the deaf and hard-of-hearing community, enabling more practical and scalable applications across diverse settings.

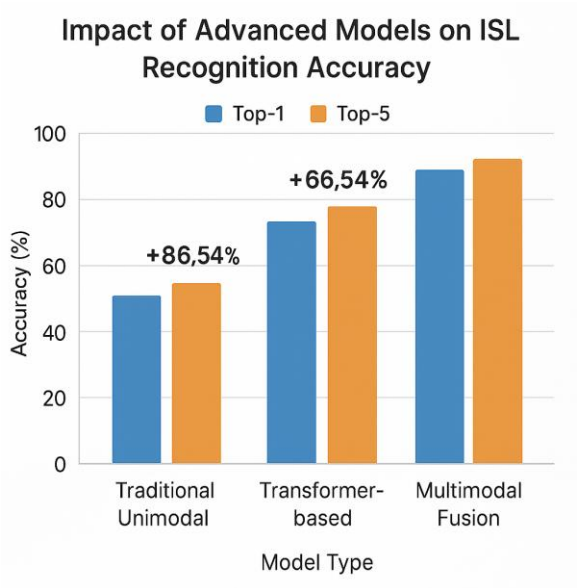

Fig. 04: Recognition Accuracy

V. COMPARATIVE ANALYSIS

While many models specialize in distinct approaches—such as handcrafted feature extraction using Support Vector Machine (SVM) and Artificial Neural Networks (ANN) [1][2], which depend on domain knowledge to manually extract key features from visual inputs, making them simpler and less computationally demanding, yet highly sensitive to variations in lighting, background, and signer-specific differences—others like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) [3][4] are specifically designed to capture both spatial and temporal patterns in gesture recognition tasks. CNNs excel at extracting visual features from individual frames, while LSTMs are capable of learning long-term dependencies across sequential movements. This architectural synergy has proven effective for recognizing continuous gestures, but the computational intensity and memory demands of CNN-LSTM models often hinder their scalability in real-time or low-resource environments, limiting their practical deployment in mobile applications.

In contrast, real-time optimization is prioritized in models like YOLO (You Only Look Once) and MobileNet [5][6], which are specifically designed to deliver high-speed inference with reduced computational loads. YOLO's single-stage detection mechanism enables it to identify hand gestures swiftly, while MobileNet's lightweight architecture is optimized for mobile and embedded systems. These qualities make them ideal for on-the-go sign language recognition; however, this efficiency frequently comes at the expense of granularity and accuracy, particularly when identifying subtle finger movements or complex hand configurations in diverse environments.

Emerging technologies such as Transformer-Based Models [7][8] represent a significant leap forward by leveraging self-attention mechanisms to understand long-range dependencies within video sequences. Unlike traditional models that struggle with temporal coherence in extended gestures, Transformers excel at capturing contextual relationships, enabling more nuanced understanding of continuous sign language. Despite their potential, these models require extensive datasets for training and substantial computational resources, posing challenges for real-time application and mobile deployment. Additionally, they often lack optimization for regional dialects and cultural variations, which can reduce their effectiveness in diverse real-world settings.

Our proposed study seeks to address these critical gaps by introducing a lightweight, real-time sign language recognition system that integrates multimodal emotion recognition, dynamic frame sampling, and culturally adaptive features within a unified, mobile-optimized framework. Unlike existing solutions, our approach not only interprets hand gestures but also incorporates facial expression cues to enhance context awareness and emotional sensitivity in recognition. This is particularly significant for improving communication accuracy in nuanced conversations, where emotions play a key role. Furthermore, the model is designed to dynamically adjust frame sampling rates based on gesture complexity, optimizing both processing speed and accuracy in real-time environments.

To tackle the prevalent issues of regional ISL variations and limited mobile deployment, our framework includes adaptive learning mechanisms that fine-tune recognition based on local dialects and device constraints, ensuring consistent performance across diverse user groups and low-resource settings. This holistic integration of gesture and emotion recognition, optimized for mobile use, represents a pioneering step forward in real-time sign language interpretation, addressing long-standing limitations in existing literature while setting a new standard for cultural and contextual sensitivity in communication technologies.

VII. CONCLUSION

The integration of deep learning and Natural Language Processing (NLP) into audio-visual translation frameworks offers a promising solution to bridge the communication gap between Indian Sign Language (ISL) users and the broader society, making ISL more accessible for the deaf community in India. This research highlights the significant advancements in ISL translation through CNN-LSTM models, transformer-based systems, and multimodal emotion recognition, which have improved translation accuracy and contextual understanding. However, challenges remain, particularly in the areas of dataset comprehensiveness, the integration of cultural and emotional context, and the optimization of these models for mobile and edge-device deployments. While some may argue that current models are sufficient, the growing demand for contextually accurate, emotion-aware translations underscores the need for further development. Future research should

focus on expanding and diversifying ISL datasets, incorporating cultural and emotional context into translation models, and optimizing dynamic sampling and multimodal fusion techniques for mobile devices. By addressing these areas, ISL translation systems can become more accurate, inclusive, and adaptable, offering better communication solutions for the deaf community in India.

REFERENCES

[1] C. Aparna and M. Geetha, "CNN and Stacked LSTM Model for Indian Sign Language Recognition," in *Machine Learning and Metaheuristics Algorithms, and Applications*, pp. 126–134, Springer, 2020.

[2] J. Joshi and D. Patel, "Transformer-Based Approach for ISL Recognition," *IJARESM*, 2023.

[3] P. Bharadwaj et al., "ISL Recognition Using YOLOv5," *IJARESM*, 2024.

[4] S. Jagtap et al., "Real-Time Sign Recognition Using MobileNetV2 and Transfer Learning," 2024.

[5] N. C. Camgöz et al., "Neural Sign Language Translation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[6] J. Huang et al., "Video-Based Sign Language Recognition without Temporal Segmentation," *arXiv preprint arXiv:1801.10111*, 2018.

[7] S. Zhang et al., "Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, pp. 3030–3043, 2018.

[8] D. Michelsanti et al., "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1368–1396, 2020.

[9] J. Huang and B. Kingsbury, "Audio-Visual Deep Learning for Noise-Robust Speech Recognition," in *ICASSP*, 2013.

[10] K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," in *IEEE Int. Conf. Big Data*, pp. 4896–4899, 2018.

[11] R. A. Calvo et al., "Natural Language Processing in Mental Health Applications Using Non-Clinical Texts," *Natural Language Engineering*, vol. 23, pp. 649–685, 2017.

[12] M. S. Hossain and G. Muhammad, "Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[13] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 1–7, 2021.

[14] K. Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," in *CVPR*, pp. 1625–1634, 2018.

[15] T. Cai et al., "Natural Language Processing Technologies in Radiology Research and Clinical Applications," *Radiographics*, vol. 36, no. 1, pp. 176–191, 2016.

[16] K. K. Dutta and S. Bellary, "Machine Learning Techniques for Indian Sign Language Recognition," in *CTCEEC*, pp. 333–336, 2017.

[17] Kim et al., "Techniques for Detecting the Start and End Points of Sign Language Utterances," [Details as provided in [45]].

[18] [Additional studies on dynamic sampling adjustments in mobile environments; see also reference [45]].

[19] J. Ekbote and M. Joshi, "Indian Sign Language Recognition Using ANN and SVM Classifiers," in *ICIIECS*, pp. 1–6, 2017.

[20] K. M. Divyashree, "Gesture Recognition for Indian Sign Language using HOG and SVM," *IRJET*, vol. 6, pp. 1697–1701, 2019.

[21] S. Katoch et al., "Indian Sign Language Recognition Using SURF with SVM and CNN," *Array*, vol. 14, p. 100141, 2022.

[22] A. Bastanfard et al., "A Novel Multimedia Educational Speech Therapy System for Hearing Impaired Children," in *PCM*, pp. 705–715, 2010.

[23] D. Kothadiya et al., "SIGNFORMER: DeepVision Transformer for Sign Language Recognition," *IEEE Access*, vol. 11, pp. 4730–4739, 2022.

[24] A. S. Dhanjal and W. Singh, "An Automatic Conversion of Punjabi Text to Indian Sign Language," *ICST Trans. Scalable Inf. Syst.*, vol. 7, pp. 1–10, 2020.

[25] R. Dhiman et al., "A Deep Learning Approach for Indian Sign Language Gestures Classification with Different Backgrounds," *J. Phys.: Conf. Ser.*, vol. 1950, pp. 1–15, 2021.

[26] S. Dhivyasri et al., "An Efficient Approach for Interpretation of Indian Sign Language Using Machine Learning," in *ICPSC*, pp. 130–133, 2021.

[27] Y. Du et al., "Full Transformer Network with Masking Future for Word-Level Sign Language Recognition," *Neurocomputing*, vol. 500, pp. 115–123, 2022.
[28] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.
[29] Kerneler, "ISL Dataset Double Handed," Kaggle, 2022.
[30] D. Kothadiya et al., "SIGNFORMER: DeepVision Transformer for Sign Language Recognition," *IEEE Access*, vol. 11, pp. 4730–4739, 2023.
[31] S. Krishna et al., "Selfie Video Based Continuous Indian Sign Language Recognition System," *Ain Shams Eng. J.*, vol. 9, pp. 1929–1939, 2018.
[32] D. Li et al., "Transferring Cross-Domain Knowledge for Video Sign Language Recognition," in *CVPR*, pp. 6205–6214, 2020.
[33] M. G. Lanjewar et al., "Fusion of Transfer Learning Models with LSTM for Detection of Breast Cancer Using Ultrasound Images," *Comput. Biol. Med.*, vol. 169, p. 107914, 2024.
[34] W. Li et al., "Sign Language Recognition Based on Computer Vision," in *ICAICA*, pp. 919–922, 2021.
[35] H. M. Mariappan and V. Gomathi, "Real-Time Recognition of Indian Sign Language," in *ICCIDS*, pp. 1–6, 2019.
[36] G. S. Mishra et al., "English Text to Indian Sign Language Machine Translation: A Rule-Based Method," *IJITEE*, vol. 8, pp. 460–467, 2019.
[37] A. Mittal et al., "A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion," *IEEE Sensors J.*, vol. 19, no. 16, pp. 7056–7063, 2019.
[38] P. Molchanov et al., "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Network," in *CVPR*, pp. 1–7, 2016.
[39] B. Natarajan et al., "Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation," *IEEE Access*, vol. 10, pp. 104358–104374, 2022.
[40] M. Oudah et al., "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques," *J. Imaging*, vol. 6, pp. 1–29, 2020.
[41] D. U. Patel and J. M. Joshi, "Deep Learning Based Static Indian-Gujarati Sign Language Gesture Recognition," *SN Comput. Sci.*, vol. 3, no. 5, 2022.
[42] A. M. Rafi et al., "Image-Based Bengali Sign Language Alphabet Recognition for Deaf and Dumb Community," in *IEEE GHTC*, pp. 1–7, 2019.
[43] G. A. Rao et al., "Deep Convolutional Neural Networks for Sign Language Recognition," in *SPACES*, pp. 194–197, 2018.
[44] S. Reshna and M. Jayaraju, "Spotting and Recognition of Hand Gesture for Indian Sign Language Recognition System with Skin Segmentation and SVM," in *WiSPNET*, pp. 1–5, 2017.
[45] T. K. Kim and B. K. Kim, "Techniques for Detecting the Start and End Points of Sign Language Utterances to Enhance Recognition Performance in Mobile Environments," *Appl. Sci.*, vol. 14, no. 20, p. 9199, 2024.
[46] J. W. Ricketts et al., "A Scoping Literature Review of Natural Language Processing Application to Safety Occurrence Reports," *Safety*, 2023.
[47] A. K. Sahoo, "Indian Sign Language Recognition Using Machine Learning Techniques," *Massy*, vol. 397, p. 2000241, 2021.
[48] A. Sharma et al., "Benchmarking Deep Neural Network Approaches for Indian Sign Language Recognition," *Neural Comput. Appl.*, vol. 33, pp. 6685–6696, 2021.
[49] D. K. Singh, "3D-CNN Based Dynamic Gesture Recognition for Indian Sign Language Modeling," *Procedia Comput. Sci.*, vol. 189, pp. 76–83, 2021.
[50] K. Yadav, L. P. Saxena, B. Ahmed, and Y. K. Krishnan, "Hand Gesture Recognition Using Improved Skin and Wrist Detection Algorithms for Indian Sign," *J. Netw. Commun. Emerg. Technol.*, vol. 9, pp. 1–7, 2019.