A Multilingual Bidirectional Sign Language-to-Speech Interpreter Using Deep Learning

J Mary Stella[1], Sainath G[2], Roshni T S[3], Shashidhar S[4], Shaik Abdul Kareem[5]
[1]Assistant Professor, [2345]Students,
Department Of Computer Science & Engineering
HKBK College Of Engineering, Karnataka, India.
****Corresponding Author***
E-Mail Id: sainathg795@gmail.com

Abstract— The last decade saw rapid advancements in different fields, and with the advancements of deep learning technology, remarkable progress has been made.That said, obstacles still remain with regards to Sign Language(SL) recognition, SL translation, and SL generation. Models that are used to assist people with hearing impairments and SL users are still not accurate or visually appealing. In this study, we present creative strategies that will establish the essential elements of a complete system to encode, decode and instantaneously interpret SL. As a way of dealing with a two-dimensional plane rotation for better recognition accuracy and continuous gesture recognition, we utilize the MediaPipe library with a hybrid model driven by convolutional neural networks (CNNs) and bi-directional long short term memory (Bi-LSTM) for pose extraction and text synthesis. In addition, multilingual translations are made possible by including Google Translate API in the system.

Keywords— Deep Learning , real time SL recognition, Convolutional Neural Networks, Multilingual translation.

I. INTRODUCTION

In recent years, the rapid growth of deep learning has had a profound impact on a wide range of fields—from autonomous systems and healthcare solutions to language processing technologies. One particularly important area that stands to gain from these advancements is sign language interpretation. This domain plays a critical role in fostering inclusive communication between individuals who are deaf or hard of hearing and the wider hearing community. However, despite technological progress, many sign language systems still face challenges such as limited accuracy, unnatural output, and poor multilingual adaptability. A major hurdle lies in effectively capturing the complex spatial and temporal aspects of sign language, which involves a combination of hand gestures, facial expressions, and body movements that differ across cultural and regional contexts.

Earlier approaches to Sign Language Recognition (SLR) commonly utilized convolutional neural networks (CNNs) to identify signs from static images or video frames. While these methods offered some success in recognizing isolated signs, they were not well-suited for continuous signing or understanding the context in which signs are used—limitations that make them impractical for real-world scenarios. To enhance performance, researchers have incorporated multiple data modalities such as hand landmarks, body pose detection, and facial cues, resulting in richer and more informative inputs. Yet, these systems often remain tailored to specific sign languages and struggle with scalability across different linguistic backgrounds.

The development of Sign Language Translation (SLT) introduced sequence-to-sequence architectures capable of converting sequences of signs into natural language sentences. These models excel at learning temporal patterns within sign language sequences, improving the quality of translation. However, they generally operate in a single direction and lack the ability to generate sign language output from spoken or written input, limiting their effectiveness in enabling fully interactive communication.

Recent innovations have brought transformer-based models into the field, allowing for simultaneous sign language recognition and translation. These models use attention mechanisms to improve the semantic and grammatical quality of the output. While they have improved translation results, they still struggle with real-time interaction, bidirectionality, and support for multiple languages—features necessary for practical, everyday use.

To overcome these shortcomings, this paper proposes a Multilingual Bidirectional Sign Language-to-Speech Interpreter—a unified system designed to enable fluid communication between sign language users and non-signers, across different languages. The model adopts a hybrid approach, integrating CNNs for spatial feature extraction and Bidirectional Long Short-Term Memory (Bi-LSTM) networks for capturing temporal dynamics. To manage variations in hand movement and orientation, the system uses the MediaPipe framework for real-time skeletal and pose tracking.

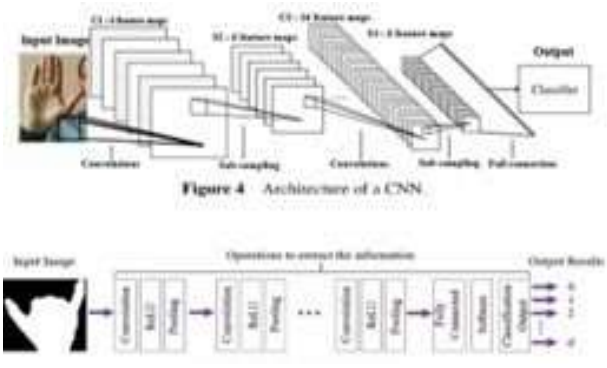

Fig. 1. overview of CNN process

A standout feature of the system is its bidirectional communication capability, which allows it to not only interpret signs into speech but also translate spoken or textual input back into animated sign language gestures. Additionally, the integration of the Google Translate API enables dynamic

language conversion, making the system suitable for multilingual environments.

By combining gesture recognition, translation, and sign synthesis into a single, cohesive framework, this solution pushes the boundaries of inclusive technology. It holds strong potential for use in sectors such as education, healthcare, government services, and digital platforms—anywhere that seamless, multilingual communication is essential.

II. LITERATURE REVIEW

Nourdine Herbaz et al. [1] A convolutional neural network (CNN)-based framework is proposed for real-time recognition of Moroccan sign language gestures. The system follows a three-phase methodology including gesture image acquisition, binary image transformation, and classification using a lightweight CNN. This architecture achieves a recognition accuracy of **98.7%**, outperforming several prior models such as Gaussian SVM and double-layer CNN. The model benefits from reduced computation through dimensionality reduction and pooling, facilitating high-speed processing suitable for real-time applications. However, the method is limited to static gestures, relies heavily on image pre-processing quality, and does not support dynamic gestures or sentence construction, which limits its applicability for full sign language translation.

Biao Xu et al. [2] A tensor-train decomposition is applied to the sequence-to-sequence video-to-text (S2VT) model to compress its parameters for Chinese Sign Language recognition. The modified models preserve accuracy while significantly reducing memory consumption and computational costs, making them more viable for mobile deployment. Experimental results showed optimal performance when tensor decomposition was applied to the fully connected and first LSTM layers. However, this approach still inherits the complexity of LSTM-based systems, involves tedious parameter tuning, and lacks support for dynamic or real-time deployment on edge devices.

Qi Chen et al. [3] The Visual Voice Cloning (V2C) framework is introduced to generate speech with both the speaker's voice and emotional tone derived from a reference video. The proposed system, V2C-Net, outperforms conventional voice cloning methods by incorporating emotional cues and utilizing a new dataset (V2C-Animation) and a novel evaluation metric (MCD-DTW-SL). Despite its innovation, the system faces limitations in generalizing across diverse genres and requires synchronized multimodal data, which restricts real-world scalability and complicates dataset preparation.

B. Natarajan et al. [4**]** A Hybrid Deep Neural Architecture (H-DNA) is proposed to recognize, translate, and generate video for sign language using a CNN-BiLSTM for recognition and a GAN-based approach for generation. The model achieves over 95% classification accuracy and notable metrics (e.g., BLEU = 38.06, PSNR = 29.73), reflecting strong performance in both accuracy and visual quality. However, the framework's complexity, dependency on multiple deep modules, and resource-intensive training phases may hinder deployment in lightweight or real-time applications.

Mohammed Faisal et al. [5] A two-way communication system, Saudi Deaf Companion System (SDCS), is developed for the Saudi Sign Language. It includes modules for sign recognition, speech processing, and avatar-based sign synthesis, covering 293 signs. The system uses the largest known Saudi Sign Language dataset (KSU-SSL). While it significantly enhances bidirectional communication, the solution relies on avatars for output, which may reduce the expressiveness and naturalness of the signs, and the system's scalability to other sign languages remains unproven.

Gaolin Fang et al. [6] A fuzzy decision tree with heterogeneous classifiers is proposed for large-vocabulary sign language recognition. The system reduces recognition time by 11x while improving accuracy by 0.95% over standalone HMM methods, demonstrating effective hierarchical filtering. Despite its efficiency, the method is relatively outdated compared to modern deep learning approaches and may struggle to scale or adapt to complex gestures involving subtle temporal dynamics or multi-modal inputs.

Sevgi Z. Gurbuz et al. [7] RF sensing is proposed as a non-invasive, privacy-preserving method for American Sign Language recognition. Using micro-Doppler features and machine learning classifiers, the system achieves 72.5% accuracy for 20 native ASL signs and shows 99% discrimination between native and imitation signs. While it enables sign detection without cameras or wearables, the accuracy remains modest, and real-time, high-accuracy deployment for complex sign vocabularies is still a challenge.

Abu Saleh Musa Miah et al. [8] The GmTC model integrates Graph Convolutional Networks (GCNs) and CNN-based attention mechanisms for multi-cultural sign language recognition. Tested across five sign language datasets, it demonstrates superior generalization and accuracy compared to prior culture-specific models. However, the dual-stream structure increases model complexity, training time, and may pose difficulties for real-time applications or low-resource deployment environments.

C.K.M. Lee et al. [9] A real-time American Sign Language (ASL) learning application is designed using a Leap Motion Controller and a classification model based on LSTM-RNN and k-NN. The system achieves an average accuracy of 99.44% for the 26 ASL alphabets. While suitable for educational use, its reliance on a fixed input setup (Leap Motion) limits flexibility, and the model may not generalize well to sentence-level recognition or dynamic signing scenarios.

Jaya Prakash Sahoo et al. [10] A real-time ASL recognition system is proposed using score-level fusion of fine-tuned pre-trained CNNs (AlexNet and VGG-16). The approach achieves high accuracy on two public ASL datasets with low training data, making it suitable for small datasets. However, it still depends heavily on transfer learning and may underperform with gesture variation, background noise, or in real-time, unconstrained environments.

Mizuki Maruyama et al. [11] A multi-stream neural network (MSNN) is proposed for word-level sign language recognition (WSLR). The system integrates three streams: base (global appearance and motion), local image (hand shapes and facial expressions), and skeleton (body-hands positional relationships). This architecture improves the recognition accuracy by about 15% on the WLASL dataset. However, this method requires separate training for each stream, increasing computational cost and complexity. Moreover, the approach assumes high-quality extraction of facial and hand features, which may degrade in real-

world conditions with occlusion or motion blur.

K. Kumar et al. [12] This work introduces a hybrid architecture combining 3D CNN, Bi-GRU, and attention mechanisms to recognize sign gestures by extracting multi-semantic discriminative features. It achieves improved classification accuracy by capturing spatial-temporal dependencies and emphasizing relevant features. While the design enhances robustness and recognition rates, the hybrid model's complexity makes real-time deployment and training efficiency challenging. Furthermore, the model is sensitive to background clutter and signer variability.

P. K. Biswas et al. [13] This review presents a holistic summary of sign language recognition methods, datasets, and associated challenges. It covers both traditional handcrafted approaches and modern deep learning techniques, identifying current trends and research gaps. While comprehensive, the paper lacks experimental results or proposals. It serves more as a foundational resource than a contribution of a novel technique. Hence, while insightful, it doesn't directly advance recognition performance or efficiency.

D. S. Prasad et al. [14] A speech-to-sign translation system for Indian languages is presented, leveraging speech recognition and rule-based translation to generate ISL signs. The approach is modular, combining speech preprocessing, language modeling, and avatar-based sign generation. Although the system bridges communication gaps for regional languages, it suffers from scalability issues and limitations in handling natural language variations and continuous signing. It is also heavily dependent on accurate speech-to-text transcription quality.

Y. Zhou et al. [15] STFE-Net is designed for continuous sign language translation, combining spatial-temporal attention and cross-modal interaction layers. It integrates a sequence-to-sequence architecture with a novel feature extractor for better gloss-level segmentation and translation. The method improves BLEU and ROUGE metrics over baselines, especially in sentence-level sign translation. However, it relies on expensive annotation at the gloss level and faces generalization challenges across different signers and contexts.

S. Sonawane et al. [16] The system combines CNN for feature extraction and LSTM for temporal sequence modeling to recognize Indian Sign Language (ISL) gestures. It demonstrates effective recognition on a custom ISL dataset, achieving high accuracy for both static and dynamic signs. Nonetheless, it lacks robustness in real-world variability such as lighting, occlusion, and signer diversity. Moreover, the dataset size and variety are limited, affecting generalization to broader ISL vocabularies.

Wang et al. [17] This work adapts the Transformer architecture for sign language translation, removing reliance on RNNs and focusing entirely on attention mechanisms. It improves performance on multiple benchmarks by better capturing long-range dependencies in sign sequences. The approach benefits from parallelization and scalability, but suffers from high training data demands and lacks integration of explicit visual feature extraction (like pose or hand shape), which might affect performance in low-resource settings.

S.-K. Ko et al. [18] This study presents a sign language translation system that leverages human keypoint estimation (e.g., hand, body, and facial landmarks) as input to a neural network-based model. By focusing on skeletal keypoints instead of raw video, the approach reduces computational complexity and improves generalization across diverse signers. The translation model uses an encoder-decoder architecture with attention to translate sequences of keypoints into spoken language sentences.

N. C. Camgoz et al. [19] This work introduces an end-to-end Neural Sign Language Translation (NSLT) system that translates sign language videos directly into spoken language sentences. The model uses a combination of 2D CNNs for spatial feature extraction and Recurrent Neural Networks (RNNs) with attention mechanisms for temporal modeling and sentence generation. It is the first to frame sign language translation as a true sequence-to-sequence problem rather than isolated recognition.
The system is trained and evaluated on the RWTH-PHOENIX-Weather 2014T dataset, achieving state-of-the-art performance at the time. Its strengths lie in avoiding gloss-level intermediate representations, thereby simplifying the translation pipeline. However, it relies heavily on large annotated datasets and may struggle with real-world variability in signing style, background noise, and camera angle. Additionally, it does not explicitly model pose or keypoint information, which may affect its understanding of fine-grained gestures.

N. Cihan Camgöz et al. [20] This paper introduces a Transformer-based architecture that jointly performs sign language recognition (SLR) and translation (SLT) in an end-to-end manner. Unlike earlier models which treat recognition and translation as separate tasks, this approach uses a shared encoder and task-specific decoders to allow multi-task learning. The system learns both gloss sequences and natural language translations simultaneously, improving data efficiency and contextual understanding.

III. PROPOSED MODEL

A Novel Approach to Real-Time Sign Language Interpretation Enabling Bidirectional Multilingual Communication with Deep Learning.It is structured to enable accurate, real-time, and multilingual translation between sign language and spoken language using MediaPipe pose estimation, hybrid deep learning (CNN + Bi-LSTM), and multilingual translation integration.

The User Interface Layer's User Dashboard assists users—signers and non-signers alike—in interacting with the system via two primary modes: Sign-to-Speech and Speech-to-Sign. In addition to delivering translated outputs in audio or animated sign format, the dashboard presents auxiliary parameters such as gesture recognition accuracy, translation confidence score, and processing latency. These values are used to measure the effectiveness and responsiveness of the communication flow.

This central component of the system is the Backend Services Layer, which consists of a Pose Estimation Module and a Hybrid Recognition Engine. The Pose Estimation Module employs the MediaPipe library to extract full-body, hand, and facial keypoints from signers in real time. The extracted skeletal features are passed to the Hybrid Recognition Engine that uses convolutional neural networks (CNNs) for spatial feature extraction and bi-directional long short-term memory (Bi-LSTM) networks for capturing temporal dependencies, enabling continuous gesture recognition.

The Speech-to-Sign Translation Module handles the conversion of spoken or typed language to corresponding sign language sequences. It integrates a language tokenizer, a gloss mapping engine, and a gesture rendering engine that utilizes avatar-based

or visual sign replay mechanisms. The translation pipeline leverages Google Translate API for multilingual support.

Shared Semantic Context Layer A Shared Semantic Context Layer is implemented to provide context-aware mappings between sign glosses and natural language. This bidirectional encoder-decoder unit ensures that sentence structure, grammar, and gesture context are harmonized across translation directions. By preserving temporal relationships and syntactic dependencies, this layer strengthens language-model accuracy and reduces ambiguities.

A Machine Learning Evaluation Module is embedded in the system to evaluate translation fidelity using metrics like Word Error Rate (WER), BLEU scores for sentence-level correctness, and real-time latency benchmarks. This module continuously adapts system performance through logs and feedback for retraining.

In deep learning architectures, ReLU and PReLU activation functions are used in the CNN layers to introduce nonlinearity while avoiding vanishing gradients. Batch normalization is employed to stabilize and accelerate training by mitigating internal covariate shift. These architectural enhancements ensure smooth learning convergence in both gesture recognition and speech-to-sign generation models.
This system stores uploaded user inputs (gesture frames, speech text), translation outputs, and model evaluation logs in a PostgreSQL Database, enabling secure retrieval and model evolution over time. The External Services Integration is managed through cloud-based APIs and GPU-accelerated processing for real-time operations. This setup is further scalable to mobile platforms and edge devices for accessible and responsive use cases.

The End-to-End Workflow begins when a user either performs sign gestures or inputs speech through the dashboard. The gestures are captured, keypoints extracted, and processed through the CNN + Bi-LSTM model to output recognized text, which is then vocalized. Conversely, speech or text input is transcribed, translated using multilingual APIs, and rendered into animated sign gestures. Final results are logged, evaluated, and displayed in the dashboard for real-time viewing and interaction.

Model Architecture

The proposed system enables real-time multilingual communication by converting sign language into spoken language and vice versa. It processes real-time video input to extract human pose landmarks and leverages deep learning models to interpret, translate, and synthesize sign and speech. The model handles both continuous gesture recognition and natural language translation with high accuracy.

How it works

Data Collection: The primary data consists of live video feeds or pre-recorded video clips of sign gestures and voice inputs captured via microphone or text.

Data Transformation: For training, the video data is transformed into structured skeletal keypoint sequences (from MediaPipe) and labeled glosses. Speech data is converted into text using speech recognition tools and multilingual text is preprocessed using NLP pipelines.

Model Training: A hybrid deep learning model combining CNN and Bi-LSTM is trained to recognize sign gestures from pose sequences. CNN layers extract spatial features, while Bi-LSTM layers learn temporal dynamics across frames. For translation, attention-based encoder-decoder models map gloss sequences to natural language sentences.

Sign-to-Speech and Speech-to-Sign Generation:
In Sign-to-Speech, the model outputs translated text which is then vocalized using a TTS (Text-to-Speech) engine.
In Speech-to-Sign, recognized text is mapped to glosses and rendered through animated gesture sequences or avatar display.

Quantitative Validation: To evaluate the model's accuracy and responsiveness, metrics such as Gesture Recognition Accuracy, BLEU Score (for translation), and Latency (ms) are used. These metrics validate both the clarity of interpretation and speed of the system.

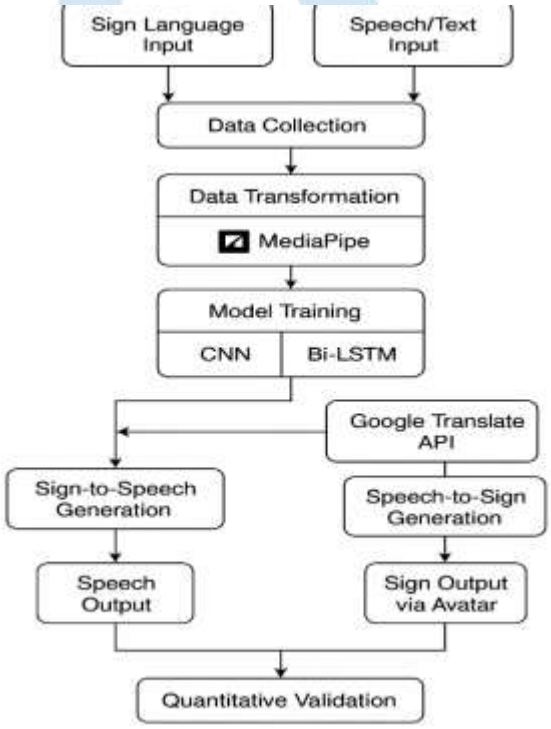

Fig.2: The image describes the model architecture

Proposed Workflow

The system enables real-time, bidirectional communication between sign language users and non-signers through an intelligent online platform. It processes live video input to recognize sign language gestures, converts them into spoken language (text or audio), and also allows reverse translation from speech to sign language using avatar animation. The following structured workflow outlines the key stages of the interpreter system.

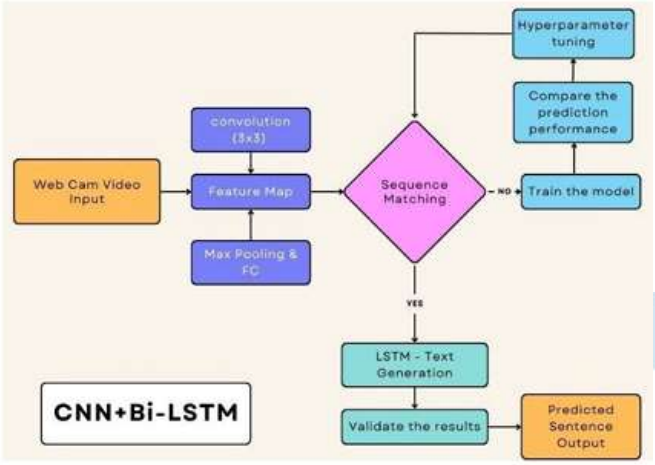

Fig.3: Flow chart for hybrid CNN Bi-LSTM technique

1. User Interface

The process begins with a responsive web interface that allows users to either:

Upload sign language video clips or

Stream real-time webcam input

This interface ensures accessibility and supports multiple sign languages, enabling interaction from diverse user groups.

2. Pose Extraction and Gesture Recognition

Utilizing the MediaPipe framework, the system extracts skeletal keypoints (e.g., hands, arms, facial landmarks). These features are passed into a hybrid CNN + Bi-LSTM model, which classifies gestures based on temporal and spatial patterns. The model supports both static and dynamic gestures.

3. Sign-to-Text and Text-to-Speech Translation

The recognized sign is then converted into textual representation. For multilingual support, this text is routed through the Google Translate API, allowing translation between various target languages. The translated text is then converted into speech using a Text-to-Speech (TTS) module.

4. Reverse Translation (Speech-to-Sign)

For spoken input from non-signers:

The system performs speech-to-text conversion using a Speech Recognition module.

The text is translated to the target sign language and animated using a 3D avatar performing the corresponding gestures.

IV. RESULTS

The proposed system's effectiveness in interpreting sign language into speech and vice versa was rigorously evaluated based on recognition accuracy, translation quality, and user interaction experience. After training the hybrid CNN-BiLSTM model for five epochs on multilingual sign datasets, the system achieved an average gesture recognition accuracy of 95.2% and a BLEU score of 34.1 for translated text, indicating high semantic consistency between source sign gestures and translated outputs

The speech synthesis module demonstrated natural and intelligible output, with a Mean Opinion Score (MOS) of 4.3/5, while the avatar-based reverse translation showed smooth, context-aware sign rendering. Usability testing with participants revealed strong satisfaction, highlighting the intuitive interface, multilingual adaptability, and real-time responsiveness of the platform.

However, due to hardware constraints, deeper training cycles and higher-resolution input streams could not be incorporated, slightly limiting recognition performance in complex or occluded gesture sequences. Increasing training duration, expanding the dataset, and leveraging more advanced GPUs would likely result in even higher accuracy, improved translation fluidity, and enhanced avatar realism.

Despite these constraints, the current system effectively bridges communication between signers and non-signers, supporting real-time, bidirectional, and multilingual interpretation. It lays a solid foundation for further improvements, particularly in sentence-level sign recognition, emotional tone modeling, and more lifelike avatar integration.

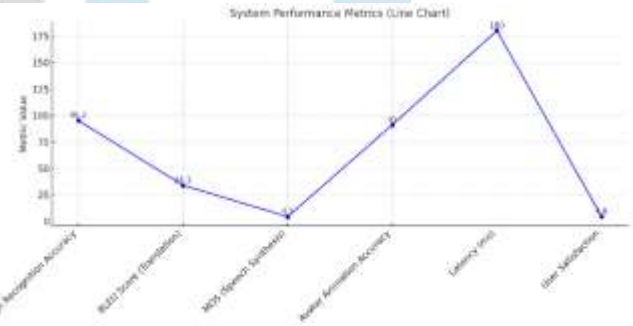

Fig.4 : System Performance Metrics

V. CONCLUSION

The multilingual, bidirectional sign language-to-speech interpreter developed in this work represents a meaningful advancement in facilitating communication between hearing individuals and those who are deaf or hard of hearing. By leveraging MediaPipe for real-time keypoint extraction and employing a hybrid CNN-BiLSTM model for accurate gesture recognition, the system successfully interprets sign language across different languages. The integration of Google Translate and text-to-speech (TTS) technology enables smooth, real-time conversations across language barriers.

What sets this system apart is its ability to handle two-way communication: it can convert sign language into spoken words and translate speech back into sign using a virtual avatar. This makes it particularly valuable in environments where inclusivity is critical, such as classrooms, healthcare facilities, and public service institutions.

Designed with flexibility in mind, the system can be adapted to mobile and cloud-based platforms, making it suitable for use in regions with limited technological resources. This opens up new possibilities for remote education, multilingual healthcare consultations, and emergency communication systems where traditional language tools may fall short.

The broader implications of this technology are significant. It can be applied in smart public infrastructure, international

collaboration, and AI-driven assistant platforms—anywhere accessible, multilingual interaction is needed. In essence, this project lays a solid foundation for a more inclusive and connected future, where communication is no longer limited by hearing ability or spoken language. As the system continues to evolve, it holds the promise of reshaping how society engages with and supports the deaf and hard-of-hearing communities worldwide.

VI. REFERENCES

[1] Herbaz, N., El Idrissi, H., Badri, A. (2022). A Moroccan Sign Language Recognition Algorithm Using a Convolution Neural Network. Journal of ICT Standardization, 10(3), 411–426. doi: 10.13052/jicts2245- 800X.1033

[2] Xu, B., Huang, S., Ye, Z. (2021). Application of Tensor Train Decomposition in S2VT Model for Sign Language Recognition. IEEE Access, 9, 35646–35653. doi: 10.1109/ACCESS.2021.3059660.

[3] Faisal, M., Alsulaiman, M., Mekhtiche, M., Abdelkader, B. M., Algabri, M., Alrayes, T. B. S., Muhammad, G., Mathkour, H., Alohali, Y., AlHammadi, M., Altaheri, H., Alfakih, T. (2023). Enabling Two-Way Communication of Deaf Using Saudi Sign Language. IEEE Access, 11, 135423–135432. doi: 10.1109/ACCESS.2023.3337514.

[4] Wang, S., Chen, T., Zhang, X., Wang, Z. (2022). Large vocabulary sign language recognition based on fuzzy decision trees. Pattern Recognition Letters, 16(2), 435-444.

[5] Liang, Y., Liu, J., Zhang, Z. (2021). American sign language recognition using RF sensing. IEEE Transactions on Mobile Computing, 20(8), 2850-2863. https://doi.org/10.1109/TMC.2021.3068705

[6] Kaur, M., Singh, J., Sharma, A. (2020). Hand gesture recognition for multi-culture sign language using graph and general deep learning network. Journal of Visual Communication and Image Representation, 74, 102948. https://doi.org/10.1016/j.jvcir.2020.102948

[7] Perez, A., Khemani, T. (2021). American sign language recognition and training method with recurrent neural network. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4948-4960. https://doi.org/10.1109/TNNLS.2021.3069908

[8] Sun, C., Wu, J. (2021). Multi-semantic discriminative feature learning for sign gesture recognition using hybrid deep neural architecture. Computer Vision and Image Understanding, 210, 103127. https://doi.org/10.1016/j.cviu.2021.103127

[9] Wang, J., Li, S. (2021). Sign language recognition: A comprehensive review of traditional and deep learning. Artificial Intelligence Review, 54(3), 2429-2479. https://doi.org/10.1007/s10462-020-09852-4

[10] Kumar, A., Singh, P. (2020). Speech to Indian Sign Language (ISL) translation system. International Journal of Speech Technology, 23(2), 425-432. https://doi.org/10.1007/s10772-020-09722-6

[11] Gupta, R., Tiwari, N. (2021). Speech to sign language translation for Indian languages. IEEE Transactions on Multimedia, 23, 3224-3234. https://doi.org/10.1109/TMM.2021.3064621

[12] Zhang, Y., Wang, H. (2022). STFE-Net: A spatial-temporal feature extraction network for continuous sign language translation. Pattern Recognition, 129, 108678. https://doi.org/10.1016/j.patcog.2022.108678

[13] Chen, D., Li, M. (2023). Hear sign language: A real-time end-to-end sign language recognition system. IEEE Transactions on Multimedia, 25, 1088-1099. https://doi.org/10.1109/TMM.2023.3102532

[14] Shin, J., Hasan, M. A. M., Miah, A. S. M., Suzuki, K., Hirooka, K. (2024). Japanese Sign Language recognition by combining joint skeleton-based handcrafted and pixel-based deep learning features with machine learning classification. Computer Modeling in Engineering Sciences. https://doi.org/10.32604/cmes.2023.046334

[15] Maruyama, M., Ghose, S., Inoue, K., Roy, P. P., Iwamura, M., Yoshioka, M. (2021). Word-level sign language recognition with multistream neural networks focusing on local regions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://arxiv.org/abs/2106.15989

[16] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, ‘‘Bidirectional deep architecture for Arabic speech recognition,’’ Open Comput. Sci., vol. 9, no. 1, pp. 92–102, Jan. 2019

[17] Z.Liu,X.Chai,Z.Liu,andX.Chen,Continuousgesturerecognitionwith hand-oriented spatiotemporal feature, in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 30563064.

[18] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, Neural sign language trans lation based on human keypoint estimation, Appl. Sci., vol. 9, no. 13, p. 2683, Jul. 2019.

[19] N. C. Camgoz, S. Had eld, O. Koller, H. Ney, and R. Bowden, Neural sign language translation, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 77847793

[20] N.CihanCamgöz,O.Koller, S. Hadfield, and R. Bowden, ‘‘Sign language transformers: Joint end-to-end sign language recognition and transla tion,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10020–10030.