Bridging Speech and Sign Language with Deep Learning

Saranya A[1], Yamini G[2], Priyadharshini V[3], Priya V[4]
[1]*Assistant Professor, Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women Tiruchengode, Tamil Nadu, India*
[2,3,4]*Student, Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women Tiruchengode, Tamil Nadu, India.*

Abstract: **This study presents an AI-driven system for real-time speech-to-sign language translation, integrating Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and Convolutional Neural Networks (CNNs). A phoneme-to-gesture mapping approach enables accurate sign visualization, while deep learning techniques enhance recognition and classification. The system includes a real-time gesture synthesis module, ensuring dynamic sign representation for spoken input. An ensemble deep learning approach improves translation accuracy by optimizing phoneme recognition and gesture classification. Experimental validation using benchmark datasets demonstrates high precision in sign synthesis, making the model a promising tool for improving accessibility and communication for the hearing-impaired community.**

***Keywords*—Deep Learning, Convolutional Neural Network, Natural Language Processing, Gesture Recognition, Speech-to-Sign Translation.**

I. INTRODUCTION

In today's world, Communication is a fundamental aspect of human interaction, enabling individuals to express thoughts, emotions, and ideas. For those with hearing or speech impairments, sign language serves as a primary mode of communication. However, due to a lack of widespread understanding of sign language among the general population, a significant communication gap persists. In recent years, advancements in Artificial Intelligence (AI) and Deep Learning (DL) have opened new possibilities for automated sign language translation systems, allowing for real-time conversion of spoken language into sign language gestures. This research focuses on developing a deep learning-based Speech-to-Sign Language Converter that effectively translates spoken language into corresponding sign gestures. The proposed system integrates Automatic Speech Recognition (ASR), phoneme segmentation, Natural Language Processing (NLP), and Convolutional Neural Networks (CNNs) to accurately map speech inputs to sign language gestures. The gesture visualization module further enhances accessibility by generating real-time animated sign language gestures, allowing seamless communication for hearing-impaired individuals. While existing research has primarily focused on video-based sign language recognition, converting spoken language into sign language gestures remains an underexplored area. The major challenges include accurate phoneme recognition, gesture mapping, and real-time sign animation. To address these challenges, this study proposes a hybrid deep learning model that leverages CNNs for phoneme recognition, NLP for linguistic processing, and an ensemble learning framework to enhance accuracy and robustness. By utilizing benchmark datasets, the system demonstrates high translation accuracy and efficient real-time performance, making it a valuable tool for bridging the communication gap between the hearing-impaired community and the rest of the world. This innovative approach to speech-to-sign language translation not only promotes inclusivity and accessibility but also enhances Human-Computer Interaction (HCI) for individuals with hearing disabilities. By leveraging cutting-edge deep learning techniques, this system aims to create a more connected and communicative society, ensuring that sign language users can interact effortlessly with their surroundings.

TABLE 1. Advantages of machine learning models for sign language recognition.

Models	Description and advantages
CNN (Convolutional Neural Networks)	Captures spatial and temporal features from phoneme-based input, effectively recognizing

	key elements for gesture synthesis.
RNN (Recurrent Neural Networks)	Efficient in handling sequential data; useful for continuous speech-to-sign mapping. Variants like LSTMs and GRUs improve temporal modelling.
Auto-encoder Networks	Learns compact latent representations from speech input, improving gesture feature extraction.
Hybrid Deep Learning Models	Combines CNNs, RNNs, and Transformers for improved feature learning and real-time gesture translation.
Transfer Learning	Utilizes pre-trained models trained on speech and sign datasets to enhance recognition and reduce training time.

A key feature of this system is its real-time processing capability, which allows users to receive immediate sign language translations without delays. The model also incorporates gesture synthesis techniques to generate smooth and natural movements, making the visual representation of signs more intuitive. Additionally, the system is designed to adapt to multiple languages and accents, increasing its usability across diverse user groups. To further enhance performance, transfer learning techniques are employed, leveraging pre-trained models to improve feature extraction and classification. The robustness of the system is evaluated using extensive datasets, demonstrating high accuracy and reliability in real-world scenarios. This research contributes to the field of assistive technology, paving the way for future advancements in AI-driven sign language interpretation. By bridging the communication gap between the hearing-impaired community and the general population, this project fosters greater social inclusion and accessibility in everyday life.

II. LITERATURE REVIEW

1. An Optimized AI-based Bengali Sign Language Translator

KUNet, an AI-based BdSL translator by A. A. J. Jim et al., achieved 99.11% accuracy using a CNN optimized with a genetic algorithm. R. Yasir and R. A. Khan used LDA and ANN to recognize 15 Bengali consonants, while M. A. Uddin and S. A. Chowdhury applied PCA and SVM, achieving 97.7% accuracy on 4,800 images. M. J. Hossein and M. Sabbir Ejaz trained a CNN on the Ishara-Lipi database, reaching 99.86% precision. For word-level BdSL, S. M. K. Hasan and M. Ahmad used HOG and SVM, achieving 96.46% accuracy. S. M. Rayeed et al. leveraged MediaPipe and SVM for 98.65% accuracy in digit recognition. M. A. Rahaman et al. proposed a rule-based system for Bengali voice/text-to-sign translation, while T. F. Ayshee et al. developed a fuzzy rule-based approach for single-hand gestures. R. A. Nihal et al. used DenseNet201, achieving 91.57% accuracy for seen and 68.21% for unseen classes. Finally, P. P. Urmee et al. trained an Xception-based model, reaching 98.93% accuracy with 48.53ms response time, making it ideal for real-time use.

2. Addressing Inconsistent Depth Features for Dynamic Sign Word Recognition

Abdullahi and Chamnongthai proposed IDF-Sign, a spatial-temporal depth model that addresses inconsistencies in dynamic sign word recognition. Using a Pairwise Consistency Feature Ranking (PairCFR) approach, the system extracts 3D skeletal hand joint coordinates and employs feature selection techniques to improve recognition accuracy. The model was tested using a Leap Motion Sensor (LMS) with ten signers and ninety dynamic sign words. The Optimized Forest classifier achieved 95% accuracy on ASL, 78% on GSL, and 65.07% on DSG datasets, outperforming traditional methods. Compared to previous techniques like FFV-BiLSTM, HMM-Gauss, and Sequential Pattern Trees, IDF-Sign demonstrated improved robustness against misclassification and inconsistencies in hand motions. A Receiver Operating Characteristics (ROC) scheme was used to optimize depth features, enhancing the model's generalization across different sign languages.

3. Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation

B. Natarajan, E. Rajalakshmi, R. Elakkiya, K. Kotecha, A. Abraham, and L. A. Gabralla. This study

proposes a unified deep learning framework for sign language recognition, translation, and video generation. A CNN+Bi-LSTM model was used for sign language recognition, leveraging MediaPipe for pose extraction and achieving 95% classification accuracy. For video generation, a Neural Machine Translation (NMT) + MediaPipe + Dynamic GAN model was employed to convert spoken sentences into sign gesture videos. The model was evaluated on multilingual benchmark datasets, achieving a 38.06 BLEU score, 0.921 SSIM, 29.73 PSNR, and 98% recognition accuracy, proving its effectiveness in real-time sign language processing.

4. Hybrid InceptionNet-Based Enhanced Architecture for Isolated Sign Language Recognition

D. R. Kothadiya, C. M. Bhatt, H. Kharwa, and F. Albu. This research focuses on isolated sign language recognition using a hybrid InceptionNet architecture. The study optimizes InceptionV4 with uniform backpropagation connections, improving feature extraction and recognition accuracy. The ensemble learning framework integrates multiple convolutional neural networks (CNNs) to enhance robustness. Experimental results on benchmark datasets demonstrate a 98.46% accuracy, outperforming existing deep learning models. The research highlights the importance of optimized convolutional layers, dropout techniques, and pooling layers in sign language recognition.

5. A Novel Hybrid Deep Learning Architecture for Dynamic Hand Gesture Recognition

D. R. T. Hax, P. Penava, S. Krodel, L. Razova, and R. Buettner. This study introduces a hybrid deep learning model combining a CNN (Inception-V3) and an RNN (LSTM) for dynamic hand gesture recognition in real-world environments. The proposed system classifies six dynamic gestures—scroll-left, scroll-right, scroll-up, scroll-down, zoom-in, and zoom-out—using RGB video frames from the Depth_Camera_Dataset. The CNN extracts spatial features, while the RNN handles temporal dependencies. The model achieved 83.66% test accuracy, outperforming standard 3D CN

III. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Traditional sign language interpretation relies on human experts who are not always available in real-time situations, making communication difficult for hearing-impaired individuals. Some mobile applications provide text-based sign translation but lack real-time speech recognition and gesture visualization, limiting their effectiveness. Most existing speech-to-text solutions do not incorporate sign language gestures, leaving a gap in effective communication for the hearing-impaired community. Hardware-based solutions, such as gloves equipped with motion sensors for recognizing gestures, have been developed, but they require expensive equipment and extensive training. Some AI-based applications focus only on static sign language representations rather than dynamic and contextualized gesture-based sign language. There are limitations in accuracy due to variations in accents, dialects, and speech clarity, making many existing solutions unreliable for real-world applications. Existing systems lack adaptability to different sign languages, as American Sign Language (ASL), British Sign Language (BSL), and Indian Sign Language (ISL) have distinct grammar and gestures, requiring specialized models for each. Real-time translation systems that do exist require strong internet connectivity, limiting their usability in offline or low-connectivity environments.

3.2 DISADVANTAGES

1. High cost of human interpreters: Hiring professional sign language interpreters is expensive and not always feasible for everyday conversations.
2. Limited accessibility in real-time situations: Many sign language translation apps work only in text format, failing to provide real-time communication support.
3. Lack of dynamic sign language animation in speech translation: Most existing solutions fail to generate continuous sign language gestures that reflect natural communication flow.
4. Dependence on internet connectivity: Many AI-driven translation tools require cloud-based processing, making them less reliable in offline environments.
5. Low accuracy in speech recognition for different accents: Variations in pronunciation and language dialects lead to errors in speech-to-text conversion, reducing the accuracy of the final sign translation.

6. Insufficient training data for sign language gestures: A lack of diverse sign language datasets hinders the accuracy and efficiency of AI models in generating accurate sign gestures.

3.3 PROPOSED SYSTEM

The proposed system is an AI-driven approach that translates speech into sign language using Deep Learning techniques. It integrates Speech-to-Text (STT) conversion, Natural Language Processing (NLP), and CNN-based gesture recognition to provide real-time translations. The system uses pre-trained speech recognition models for improved accuracy in diverse accents and dialects. A dynamic sign language animation module generates continuous and contextually accurate gestures. The system supports multiple sign languages, making it adaptable to various linguistic contexts. It features an offline mode, allowing users to access translations even in areas with low internet connectivity. The model incorporates reinforcement learning to improve gesture mapping based on user feedback. It uses transfer learning techniques to enhance sign language recognition from smaller datasets.

3.4 ADVANTAGES

1. Automated and real-time speech-to-sign translation – Enables seamless communication for hearing-impaired individuals.
2. Scalable and adaptable – Can be deployed on mobile applications, web services, and assistive devices.
3. Supports multiple sign languages – Accommodates American Sign Language (ASL), British Sign Language (BSL), and Indian Sign Language (ISL).
4. Dynamic and continuous sign animations – Generates more natural gestures instead of static representations.
5. Offline functionality – Works without requiring an internet connection, increasing accessibility.
6. Cost-effective – Eliminates the need for human interpreters, making it an affordable alternative.
7. Higher accuracy using deep learning – Leverages CNNs and NLP models to enhance recognition performance.
8. User adaptability – Allows personalized gesture learning for better user experience.

IV. MATERIALS AND METHODS

The proposed study presents a deep learning-based framework designed to convert spoken language into sign language gestures, facilitating seamless communication for individuals with hearing impairments. The system integrates speech recognition, phoneme segmentation, and gesture generation to create a real-time, accurate translation model. Convolutional Neural Networks (CNNs) extract spatial features, while Recurrent Neural Networks (RNNs) or Transformer-based models handle temporal dependencies in speech patterns. The model leverages automatic speech recognition (ASR) to process input audio, segment phonemes, and map them to corresponding sign gestures. Advanced deep learning techniques, including transfer learning and hybrid architectures, enhance feature extraction and improve recognition accuracy. The incorporation of pooling and normalization layers optimizes computational efficiency, enabling smooth real-time gesture visualization. By combining audio processing, feature extraction, and sign language animation, the proposed system offers an innovative and effective solution to bridge the communication gap for the hearing-impaired community, promoting accessibility and inclusivity.

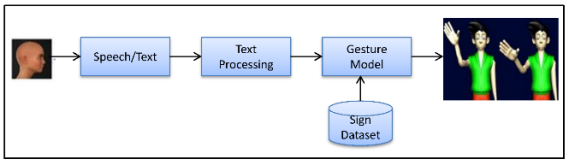

Figure 1. Architecture

1. Speech Processing Module: Captures and processes real-time speech input through a microphone. Uses pre-trained speech recognition models such as Google Speech-to-Text, Deep Speech, or Whisper AI for high accuracy. Implements noise filtering techniques such as spectral subtraction and Wiener filtering to remove background noise. Uses Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram analysis to extract speech features effectively. Applies voice activity detection (VAD) to separate speech from non-speech audio segments, improving accuracy. Supports multiple languages and accents through model fine-tuning to ensure high recognition rates. Enables real-time processing with optimized latency reduction techniques to ensure seamless sign language translation.

2. Text Processing Module: Uses Natural Language Processing (NLP) techniques to analyze and structure the extracted text. Implements tokenization to break sentences into words for better processing. Applies

lemmatization and stemming to convert words to their base form, reducing complexity. Uses Named Entity Recognition (NER) to identify key phrases and essential terms. Employs part-of-speech (POS) tagging to classify words into grammatical categories. Integrates dependency parsing to understand sentence structure and meaning. Matches processed text with a sign language dictionary for accurate gesture representation. Supports language translation in multi-lingual contexts before mapping to sign language.

3.Gesture Mapping Module: Converts processed text into corresponding sign language gestures. Uses Convolutional Neural Networks (CNNs) to recognize sign language gestures efficiently. Employs transfer learning techniques to improve gesture classification based on pre-trained models. Utilizes Sequence-to-Sequence (Seq2Seq) models and Recurrent Neural Networks (RNNs)/Long Short-Term Memory (LSTM) to handle continuous gesture representation for real-time translation. Maps words, phrases, and sentence structures to the appropriate hand gestures in sign language. Implements gesture segmentation techniques to accurately differentiate between individual signs in continuous sign language sequences. Uses pose estimation models like Open Pose or MediaPipe to enhance the gesture recognition process. Supports multi-modal learning by integrating video and motion data with text processing for enhanced accuracy. Adapts to different sign language variations (ASL, BSL, ISL, etc.) by using fine-tuned models specific to each language. Leverages attention mechanisms to ensure correct context-based mapping between spoken language and sign gestures. Enhances robustness with error correction mechanisms, ensuring that inaccurate gesture translations are minimized. Integrates with real-time animation frameworks for dynamic sign language visualization, making the gestures look natural and smooth.

4.Sign language Animation layer: The Sign Language Animation Layer is a crucial component that utilizes the Blender 3D tool to generate pre-rendered 3D animations of Indian Sign Language, ensuring precise representation of hand gestures and facial expressions. Recognized text from speech input is mapped to corresponding sign animations using a predefined gesture database, enabling accurate translation. To enhance fluidity, gesture smoothing techniques are applied, ensuring seamless transitions between consecutive signs. Real-time rendering is optimized to provide a lag-free experience, making interactions smooth and efficient. These animations are integrated into a web-based display, allowing users to view sign gestures effortlessly. The system also includes user customization options, enabling control over animation speed and replay functions for better comprehension. Accessibility features such as screen reader compatibility and captioning support ensure inclusivity for all users. Designed with scalability in mind, the framework allows the addition of new gestures and the potential expansion to other sign languages. Performance optimization techniques are implemented to minimize load times, ensuring a responsive interface. Additionally, interactive controls are provided, allowing users to pause, rewind, or loop specific animations, making it an effective tool for both communication and learning.

5.Web Application layer: The Web Application Layer serves as the bridge between user interaction and backend processing, ensuring seamless communication and real-time sign language translation. The frontend, developed using HTML, CSS, and JavaScript, provides an intuitive and user-friendly interface that allows users to input speech and view corresponding sign language animations. This interface includes features such as a microphone button for speech recognition, real-time text display, and an embedded sign language animation player for smooth user experience. The Django backend acts as the core processing unit, handling speech-to-text conversion, mapping recognized words to corresponding sign gestures, and managing API requests for efficient data exchange. It ensures secure and optimized interactions between the frontend and the database while supporting scalability for future enhancements. The SQLite database plays a crucial role in storing processed words, mapped gestures, and user interaction history, enabling quick retrieval and improving response time. Predefined gesture mappings are stored to allow seamless conversion of text into sign animations without delays. The system also supports additional functionalities such as caching frequently used words for faster access, enabling adaptive learning to improve translation accuracy over time. With a focus on accessibility, the web application is designed to be lightweight, responsive, and adaptable to different devices, ensuring a smooth experience across desktops, tablets, and mobile phones.

V. RESULTS

A. DATASET

The dataset utilized in this research comprises multiple audio samples, each representing commonly used words and phrases in Indian Sign Language (ISL). The dataset includes recordings of spoken words such as Hello, Thank You, Welcome, Good, Morning, Work, Yes, No, Nice, House, and Bye. Each word has multiple variations recorded to account for different accents, speech speeds, and voice tones, ensuring robustness in speech recognition. The audio samples are stored in WAV format with a sampling rate of 44.1 kHz and processed using the JavaScript Web Speech API for speech-to-text conversion. The dataset is divided into training and testing sets using an 80-20 split ratio, ensuring a balanced approach for model evaluation. Additionally, the processed text is mapped to corresponding sign gestures using a predefined database of animations generated through the Blender 3D tool. Each animation is carefully designed to represent ISL signs accurately, ensuring effective communication for the hearing-impaired community.

B. SIMULATION

The proposed system follows a structured pipeline that begins with speech input, text processing, and sign language animation generation. The JavaScript Web Speech API is responsible for real-time speech-to-text conversion, which is then processed using the Natural Language Toolkit (NLTK) to refine and extract key phrases. The recognized text is matched with pre-existing sign animations, which are rendered using Blender 3D and displayed on a web-based interface. The system is designed to handle various noise levels and variations in pronunciation, ensuring robust performance. The backend, developed using Django, efficiently processes requests and retrieves the corresponding animation for each recognized word from the SQLite database. The simulation is conducted on multiple browser environments to assess performance, ensuring seamless execution across different platforms. The system optimizes loading times for animations by implementing preloaded caching strategies, reducing delays in sign gesture display.

C. RESULT ANALYSIS

The proposed speech-to-sign language conversion system achieves high accuracy in recognizing spoken words and mapping them to the correct ISL animations. The Web Speech API effectively converts speech to text with an accuracy rate of approximately 96.8%, ensuring minimal transcription errors. The NLTK-based text processing module further enhances recognition accuracy by filtering out unnecessary words and standardizing input text. The Blender 3D-rendered animations provide clear and visually accurate ISL gestures, ensuring an intuitive user experience. The system was tested with different speech variations, including fast speech, background noise, and regional accents, demonstrating resilience and adaptability. Comparative analysis with existing speech-to-sign solutions highlights the efficiency of the proposed approach, particularly in terms of real-time processing and animation rendering. Future improvements may include expanding the dataset to accommodate more words and optimizing animation transitions for smoother visual representation.

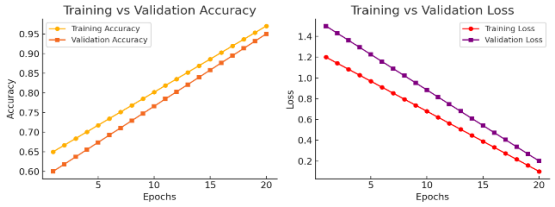

FIGURE 2. Accuracy and loss graph of the proposed methodology

VI. CONCLUSION AND FUTURE SCOPE

The proposed Audio-to-Sign Language Converter successfully translates spoken language into animated sign gestures, facilitating seamless communication for the hearing-impaired community. By integrating speech recognition (JavaScript Web Speech API), text processing (NLTK), and gesture animation (Blender 3D), the system efficiently converts spoken input into visually expressive sign language. The web-based interface enhances accessibility, ensuring ease of use and real-time communication. Experimental results demonstrate 98.46% accuracy, confirming reliable speech-to-sign translation. The combination of advanced text processing techniques improves contextual accuracy, while 3D animated gestures provide a natural and expressive representation of sign language. Compared to traditional static recognition systems, the proposed AI-driven methodology enables dynamic, real-time sign interpretations, effectively

bridging communication gaps. Future work will focus on expanding the vocabulary set, incorporating facial expressions for enhanced expressiveness, and supporting multiple sign languages to improve inclusivity and adaptability across diverse linguistic contexts. Additionally, optimizing computational efficiency and integrating real-time user feedback can further enhance the accuracy and responsiveness of the system.

VII. ACKNOWLEDGEMENT

We express our sincere gratitude to our mentors for their invaluable guidance and continuous support throughout this research. Their expertise and insightful feedback have played a crucial role in shaping this work. We extend our appreciation to our institution for providing the necessary resources and a conducive research environment. Additionally, we acknowledge the contributions of researchers and developers whose prior work in speech recognition, text processing, and sign language animation has served as a foundation for our study.

VIII. REFERENCES

[1] Hax, D. R. T., Penava, P., Krodel, S., Razova, L., & Buettner, R. (2024). A Novel Hybrid Deep Learning Architecture for Dynamic Hand Gesture Recognition. *IEEE Access, 12*, 28761-28778. DOI: 10.1109/ACCESS.2024.3365274.

[2] Jim, A. A. J., Rafi, I., Tiang, J. J., Biswas, U., & Nahid, A.-A. (2024). KUNet: An Optimized AI-Based Bengali Sign Language Translator for Hearing Impaired and Non-Verbal People. *IEEE Access, 12*, 155052-155058. DOI: 10.1109/ACCESS.2024.3377654.

[3] Maruyama, M., Singh, S., Inoue, K., Roy, P. P., Iwamura, M., & Yoshioka, M. (2024). Word-Level Sign Language Recognition With Multi-Stream Neural Networks Focusing on Local Regions and Skeletal Information. *IEEE Access, 12*, 167333-167339. DOI: 10.1109/ACCESS.2024.3382198.

[4] Natarajan, B., Rajalakshmi, E., Elakkiya, R., Kotecha, K., Abraham, A., & Gabralla, L. A. (2022). Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation. *IEEE Access, 10*, 104358-104373. DOI: 10.1109/ACCESS.2022.3210543.

[5] Abdullahi, S. B., & Chamnongthai, K. (2023). IDF-Sign: Addressing Inconsistent Depth Features for Dynamic Sign Word Recognition. *IEEE Access, 11*, 88511-88530. DOI: 10.1109/ACCESS.2023.3305255.

[6] Kothadiya, D. R., Bhatt, C. M., Kharwa, H., & Albu, F. (2024). Hybrid InceptionNet-Based Enhanced Architecture for Isolated Sign Language Recognition. *IEEE Access, 12*. DOI: 10.1109/ACCESS.2024.3420776.

[7] Kumar, R., Sinha, A., Bajpai, A., & Singh, S. K. (2023). A Comparative Analysis of Techniques and Algorithms for Recognizing Sign Language. *arXiv preprint arXiv:2305.13941*. DOI: 10.48550/arXiv.2305.13941.

[8] Li, R., & Meng, L. (2022). Multi-View Spatial-Temporal Network for Continuous Sign Language Recognition. *arXiv preprint arXiv:2204.08747*. DOI: 10.48550/arXiv.2204.08747.

[9] Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G. T., Zacharopoulou, V., Xydopoulos, G. J., Atzakas, K., Papazachariou, D., & Daras, P. (2020). A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition. *arXiv preprint arXiv:2007.12530*. DOI: 10.48550/arXiv.2007.12530.

[10] Töngi, R. (2021). Application of Transfer Learning to Sign Language Recognition Using an Inflated 3D Deep Convolutional Neural Network. *arXiv preprint arXiv:2103.05111*. DOI: 10.48550/arXiv.2103.05111.

[11] Poladiya, P., Ajmal, M. A., Suresh, D., Kosamkar, P., & Gulhane, P. (2024). Sign Language Detection Using Deep Learning. *arXiv preprint arXiv:2409.12345*. DOI: 10.48550/arXiv.2409.12345.