International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 03 | Mar 2024 www.irjet.net p-ISSN: 2395-0072

Capturing Non-manual features of Indian Sign Language and Converting it into Text

Mamatha S [1], Nithya T M [2], Prachi [3] , Narmada Radhika J S [4] , Dr.Leelambika [5]

[1234] UG student, Dept. of Computer Science & Technology, Presidency University, Bengaluru.
[5]Assistant Professor, Senior Scale, Dept. of Computer Science & Engineering , Presidency University, Bengaluru.

Abstract - *The project "Speech and Sign Language Translation to Text for Enhanced Communication" aims to bridge the communication gap between the deaf and hard-of-hearing individuals and the general public by utilizing artificial intelligence and machine learning techniques. The project leverages the American Sign Language (ASL) dataset and the YOLO v8 model to recognize live sign language gestures performed by the user and convert them into corresponding text, enabling real-time communication. Additionally, the project includes a speech-to-text conversion module that allows users to speak, and the system transcribes their speech into text using speech recognition libraries. The entire system is developed using Flask, offering an efficient and user-friendly platform. By integrating both sign language and speech recognition, this project promotes enhanced communication, fostering inclusivity and accessibility. The combined functionalities of sign language translation and speech-to-text ensure that both visual and auditory communication needs are addressed, making the system suitable for a wide range of applications in real-time communication. Furthermore, the predicted words can be converted into Kannada language, ensuring that users in regions where Kannada is spoken can also benefit from the system. Using a pretrained model, e we are recognizing these gestures: 'okay', 'peace', 'thumbs up', 'thumbs down', 'call me', 'stop', 'rock', 'live long', 'fist'.*

Key Words: Speech-to-Text, Sign Language Recognition, YOLO v8, ASL Dataset, Flask, Deep Learning, Real-Time Translation, Accessibility, Communication Enhancement, AI-based System

1.INTRODUCTION

Communication plays a pivotal role in everyday life, yet many individuals face barriers due to hearing or speech impairments. Traditional approaches often fail to provide inclusive solutions that cater to both sign language users and those requiring speech-based assistance. To address this gap, the project titled “Speech and Sign Language Translation to Text for Enhanced Communication” integrates cutting-edge deep learning methods and modern web technologies. By focusing on American Sign Language (ASL) and real-time speech recognition, it aims to create a unified platform that bridges multiple modes of communication. At the heart of this system is the YOLO v8 model, a powerful object detection algorithm capable of recognizing hand gestures within live video streams. Trained on a curated ASL dataset, YOLO v8 identifies various signs in real time, converting them to text so that users can form words and sentences. This streamlined process ensures accurate detection, even under challenging conditions, thereby making sign language more accessible to those unfamiliar with it. Additionally, the project employs a robust speech recognition library for converting spoken language into text. Whether users rely on speech or sign language, the system guarantees immediate transcription, fostering smoother interactions between diverse audiences. One of the core principles guiding this project is user-friendliness. To that end, a Flask-based web interface was chosen for its simplicity and flexibility. This allows for seamless integration of both modules—sign language recognition and speech-to-text conversion—within a single, cohesive environment. The Flask framework also makes it easy to scale and adapt the system, paving the way for future expansions such as multi-language support or enhanced models capable of recognizing subtle hand movements and different signing styles. Beyond its core functionality, this project sets an important precedent for developing inclusive communication tools. By offering both sign-to-text and speech-to-text in a unified platform, it addresses the needs of people with varying communication preferences. Educational institutions can utilize it to support deaf or hard-of-hearing students, healthcare providers can use it to improve patient engagement, and customer service platforms can deploy it to serve wider audiences. In essence, the “Speech and Sign Language Translation to Text for Enhanced Communication” project serves as a steppingstone to a more accessible digital world. By integrating YOLO v8, speech recognition libraries, and a Flask-based interface, it champions inclusive design and lays the groundwork for further innovation in real-time, multi-modal communication systems.

1.1 Statement About the Problem

Communication barriers faced by individuals with hearing impairments and those who rely on sign language pose significant challenges in daily interactions. Although sign language is a valuable means of communication, it is not widely understood by the general population, limiting accessibility and interaction for deaf or hard-of-hearing individuals. Similarly, while speech recognition technology has advanced, people with speech impairments or those who prefer non-verbal communication may face difficulties when

interacting with voice-based systems. The lack of seamless, real-time solutions that cater to both speech and sign language limits the inclusivity of communication systems. This project aims to address these issues by developing an integrated system that allows real-time translation of sign language into text using the ASL dataset and YOLO v8, while also providing a speech-to-text conversion module. The goal is to create an accessible platform for both sign language users and those with speech impairments, enhancing communication in diverse settings.

1.2 Scope of the Project

The scope of the project "Speech and Sign Language Translation to Text for Enhanced Communication" encompasses the development of a comprehensive solution to facilitate communication for individuals with hearing and speech impairments. The primary goal is to create an interactive platform that provides real-time translation of sign language gestures into text using the ASL dataset and YOLO v8, ensuring accessibility for deaf and hard-of-hearing individuals. Additionally, the system includes a speech-to-text module that converts spoken language into text, allowing users with speech disabilities or those preferring non-verbal communication to interact effectively. The project aims to deliver an easy-to-use interface built with Flask, ensuring smooth integration of both modules. The solution has the potential for broader applications in various domains, such as education, healthcare, customer service, and public spaces, promoting inclusivity and enabling seamless communication between diverse groups. The project can be expanded to support multiple languages and real-time translation in the future.

1.3 Hybrid Detection and classification model

Hybrid YOLO-CNN-RNN Architecture

In the original system, YOLO v8 (You Only Look Once version 8) was used solely for detecting hands in real-time webcam streams. While YOLO excels in fast and accurate object detection, a hybrid approach enhances its capabilities by combining it with deep learning models that specialize in sequence learning and fine-grained gesture recognition.

In the **proposed hybrid model**, YOLO v8 is utilized as the first stage of detection to locate and isolate the hand region from the video stream efficiently. This ensures that only relevant spatial data is passed to the next stages, minimizing noise from the background and improving processing speed.

Once the hand is detected, the cropped image sequence is passed to a **CNN-RNN pipeline**:

- The **CNN (Convolutional Neural Network)** extracts spatial features from each hand image frame, identifying patterns like finger placement, hand contours, and movement dynamics.
- These features are then fed into an **RNN (Recurrent Neural Network)**—specifically, a Long Short-Term Memory (LSTM) network—which captures **temporal dependencies** across a sequence of frames. This is especially beneficial for recognizing continuous gestures or signs that unfold over time.

By integrating YOLO with CNN and RNN models, the hybrid architecture delivers superior accuracy in recognizing complex signs, including dynamic movements that traditional CNNs may misclassify when processed in isolation.

This modular pipeline enhances the model's robustness and adaptability across various users, lighting conditions, and signing styles, while preserving the real-time responsiveness of the application. Moreover, this architecture opens the door for integrating advanced temporal context-aware models in future iterations, such as Transformer-based encoders for gesture translation.

2. LITERATURE SURVEY

Recent advancements in sign language recognition have leveraged a variety of sensor technologies, machine learning models, and computer vision approaches to enable more effective communication for individuals with hearing or speech impairments. Researchers have explored both static and dynamic gesture recognition using deep learning, particularly Convolutional Neural Networks (CNNs), hybrid models, and spatio-temporal processing techniques.

Vashisth et al. [1] proposed a CNN-based model tailored for Indian Sign Language (ISL) recognition. Using a custom image dataset, their model achieved a recognition accuracy of 99% with a training loss of just 0.0178. The system's high performance highlights the efficacy of CNNs in recognizing static hand gestures within ISL.

In another study, Areeb et al. [2] developed a deep learning solution for recognizing emergency-related signs to assist hearing-impaired individuals during critical scenarios. They implemented three different models, including a 3D CNN, a VGG16-LSTM hybrid, and a YOLOv5-based gesture detector. Among these, the YOLOv5 model yielded the best results with a mean average precision (mAP) of 99.6%, showcasing the potential of real-time object detection techniques in dynamic gesture applications.

Kothadiya et al. [3] introduced a hybrid architecture based on InceptionNet for isolated sign recognition. The enhanced

network architecture integrated ensemble learning and backpropagation optimization, achieving a classification accuracy of 98.46% across standard datasets. This approach proved highly robust for recognizing isolated signs with greater precision.

Addressing the inconsistencies often found in depth-based gesture data, Abdullahi and Chamnongthai [4] presented the IDF-Sign framework. Utilizing 3D skeletal joint coordinates obtained from a Leap Motion Sensor, they introduced a Pairwise Consistency Feature Ranking (PairCFR) algorithm to select stable features for dynamic sign word recognition. Their method achieved up to 95% accuracy and showed notable improvements in depth-based gesture analysis.

Rajalakshmi et al. [5] proposed a hybrid deep neural network that combines a 3D CNN with attention-based Bi-LSTM layers for multilingual sign gesture recognition. Their model, evaluated on a newly developed Indo-Russian dataset, demonstrated strong performance in extracting both spatial and temporal features, surpassing several existing methods in classification accuracy and generalization.

Collectively, these studies demonstrate significant advancements in sign language recognition through deep learning, particularly in hybrid and multi-modal systems. Nonetheless, key challenges persist, including signer variability, continuous gesture interpretation, and real-time deployment. Future work should focus on improving model adaptability, expanding dataset diversity, and optimizing lightweight architectures for embedded applications.

3. PROPOSED METHODOLOGY

3.1 Data Collection and Preprocessing

- The project begins with gathering an **American Sign Language (ASL) dataset**, including a range of hand gestures representing letters or words.
- Preprocessing steps involve **labelling and annotating** images to ensure the YOLO v8 model can accurately detect and classify gestures.
- Additional data augmentation techniques, such as **rotation, flipping, and cropping**, help make the model more robust to variations in signing styles and lighting conditions.

3.2 YOLO v8 Model Training

- The labelled ASL images are fed into the **YOLO v8** framework, which processes each image to learn features associated with specific signs.
- During training, **loss functions** track detection accuracy, guiding the model to refine its predictions with each epoch.
- Hyperparameters (e.g., learning rate, batch size) are fine-tuned to strike a balance between accuracy and computational efficiency.

3.3 Speech Recognition Module

- For speech-to-text, a **speech recognition library** (e.g., Google Speech Recognition API or CMU Sphinx) is integrated.
- Users' spoken words are **captured via microphone** input, converted into text strings, and displayed alongside the sign language output.

3.4 Integration with Flask

- A **Flask**-based web application serves as the user interface, streamlining both modules (sign recognition and speech-to-text) into one cohesive platform.
- Flask routes manage data flow, handling both **camera inputs** for sign detection and **microphone inputs** for speech recognition.

3.5 Technologies

3.5.1 Flask (Python Web Framework)

Flask is a lightweight and powerful Python web framework that is used in this project to build the web application. It provides a simple way to define routes, manage HTTP requests, render HTML templates, and integrate backend machine learning models with the frontend interface. In this project, Flask handles important functionalities like user registration, login authentication, webcam feed streaming, and audio transcription through various routes. It acts as the bridge between the users and the machine learning models, making the entire system interactive, web-accessible, and user-friendly.

3.5.2 YOLO v8 (Ultralytics Object Detection)

YOLO v8 (You Only Look Once version 8) is a cutting-edge deep learning model specialized in object detection. In this project, YOLO v8 is utilized to detect the presence and location of hands in real-time webcam video feeds. The model quickly identifies the hand region with high precision, which is then extracted and sent for gesture classification. Using YOLO v8 ensures that only the relevant part (the hand) is processed, greatly improving the accuracy and speed of the gesture recognition system. Its ability to perform

detection at high speed while maintaining excellent accuracy makes it ideal for real-time applications like this.

3.5.3 CNN (Convolutional Neural Network)

A Convolutional Neural Network (CNN) is used in this project to recognize and classify different hand gestures after the hand has been detected. CNNs are particularly powerful in learning features from image data automatically without manual feature engineering. Here, the CNN model processes the cropped hand image and predicts the specific gesture being shown, such as "Hello," "Yes," or "No." By using Keras and TensorFlow libraries, the CNN is trained and later loaded to perform fast and reliable gesture classification, ensuring that the system achieves high accuracy and robustness even in challenging conditions.

3.5.4 OpenCV (Open Source Computer Vision Library)

OpenCV is a versatile library for computer vision tasks and plays a critical role in this project by handling the webcam operations. It captures real-time video frames from the camera, processes the images (such as flipping and cropping), and displays the predicted output. OpenCV also supports drawing bounding boxes and text over frames, which helps in visualizing the detection and recognition results. Without OpenCV, integrating live camera input and interacting dynamically with image data would be highly complex.

3.5.5 MediaPipe (Hand Landmark Detection)

MediaPipe, developed by Google, is used in the project for detecting hand landmarks — specifically, the 21 key points of a human hand such as fingertips and joints. By accurately identifying these landmarks, MediaPipe provides structured information about the hand's position and pose, which can further enhance gesture recognition. Although YOLO detects the overall hand, MediaPipe allows a deeper understanding of the hand's shape and movement if required. Its real-time processing and lightweight design make it ideal for enhancing hand-tracking accuracy.

3.5.6 Speech Recognition (Audio to Text Conversion)

Speech Recognition is a Python library used to convert spoken words into text. In this project, it enables users to input information via their microphone, which is then transcribed into text. This is particularly useful in the /mic route of the application, where audio recordings are processed through Google's speech recognition API to return a transcript. Integrating Speech Recognition adds a multi-modal interaction capability to the project, allowing users to interact through both gestures and voice commands.

3.6 Model Architecture

The application begins with **system initialization**, where all essential components such as the user interface, YOLO-based detection models, and MediaPipe hand tracking tools are loaded and prepared. Once initialized, the user proceeds to **authenticate** by registering or logging in, which ensures secure access and enables personalized user interactions. Upon successful login, the user is directed to the **dashboard interface**, where they can choose their preferred **input modality**—either speech or sign language.

If the user selects the **speech input** route, the application activates the **audio signal processing module**, capturing real-time input from the microphone. This audio input is then processed by an **automatic speech recognition (ASR)** system—typically powered by cloud-based services like the Google Speech API—to transcribe spoken words into text.

Alternatively, if the user opts for **sign language input**, the application engages the **visual gesture processing module**, utilizing the webcam to monitor hand movements. The system first employs a YOLO-based object detection model to identify and isolate the hand region within each frame. It then leverages **MediaPipe** to extract detailed **hand landmarks**, and a pre-trained **Keras model** classifies these gestures into corresponding letters or signs, enabling **sign language recognition via deep learning**.

After either input type is processed, the recognized content is presented through **textual output rendering**, displaying the converted text on the user interface. Finally, when the interaction concludes, the system proceeds to **session termination**, releasing all resources and securely closing the application.

This workflow ensures a seamless and intelligent communication bridge between speech and sign language using advanced deep learning and computer vision technologies.

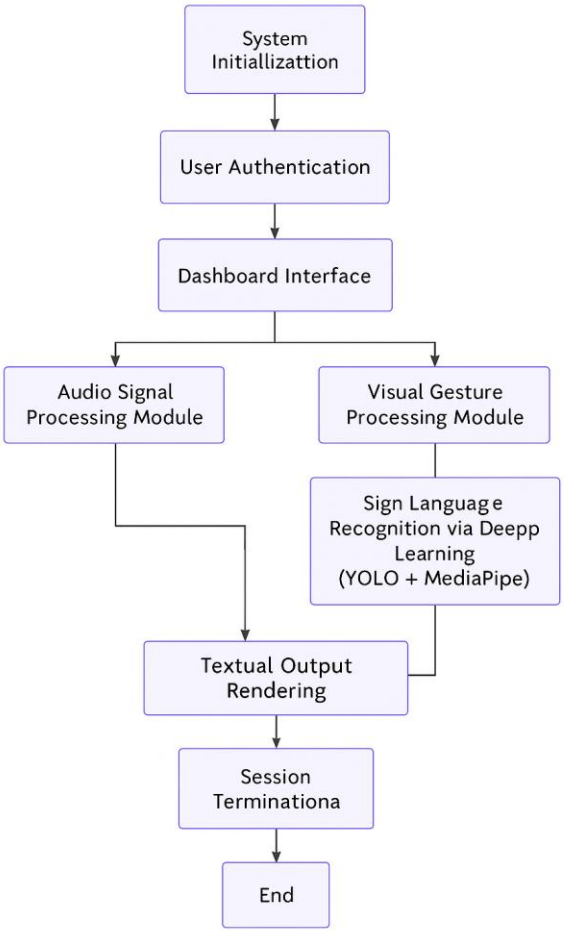

FIG.1: Model Architecture

4. RESULTS

These YOLO model results showcase an overall high level of performance in detecting and classifying various letters (A–Z) in American Sign Language (ASL). With an **average precision (Box P)** of 0.911 and **recall (R)** of 0.865, the model demonstrates reliable detection capabilities across diverse gestures. The **mAP (mean Average Precision) at 50% IoU** is notably high at **0.974**, while **mAP at 50-95%** (a more stringent metric) stands at **0.925**, indicating robust performance even under stricter evaluation conditions. Looking at per-class details, most letters, such as **C, E, H, I, J, Q, R, W**, and **Z**, achieve near-perfect precision or recall, reflecting the model's ability to accurately detect these gestures. Some classes, including **B, O, P, and T**, show slightly lower recall (ranging from 0.53 to 0.83), but still maintain strong overall performance. This suggests that while the model generally handles most letters well, specific classes may need further fine-tuning or data augmentation to improve consistency. Overall, these metrics underscore the model's suitability for real-time sign language detection tasks. However, continuous refinement—such as adding more training samples, diverse lighting conditions, and motion variations—could bolster accuracy and recall for the less robust classes.

5. CONCLUSIONS

The project **"Speech and Sign Language Translation to Text for Enhanced Communication"** successfully integrates advanced deep learning and speech recognition techniques to bridge the communication gap for individuals with hearing and speech impairments. By leveraging the **YOLO v8 model** for real-time sign language recognition and **speech recognition libraries** for speech-to-text conversion, the system provides an efficient, accessible, and inclusive solution for diverse communication needs. The use of Flask ensures seamless integration of both modules, allowing users to interact with the system effortlessly.

This project addresses the limitations of traditional sign language and speech recognition systems by improving accuracy, real-time processing, and adaptability. With potential applications in **education, healthcare, public services, and customer support**, it can significantly enhance accessibility for the deaf, hard-of-hearing, and speech-impaired individuals.

Future enhancements may include **multi-language support, gesture refinement using advanced neural networks, and integration with wearable devices** to further improve usability. This project is a step toward a more inclusive digital world, promoting seamless and effective communication for all.

OUTPUT:

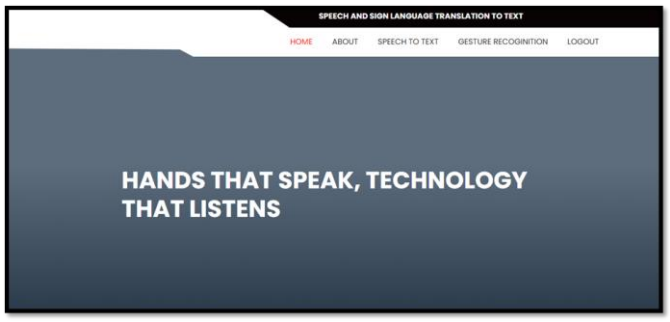

Fig.1: Application Home Page

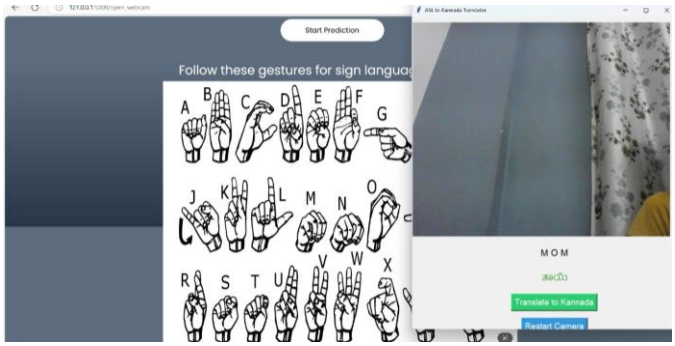

Fig.2: Kannada Translation from Sign Language

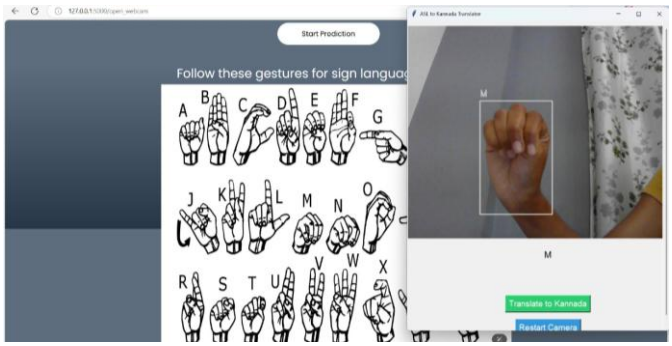

Fig.3: ASL Alphabet Reference Chart

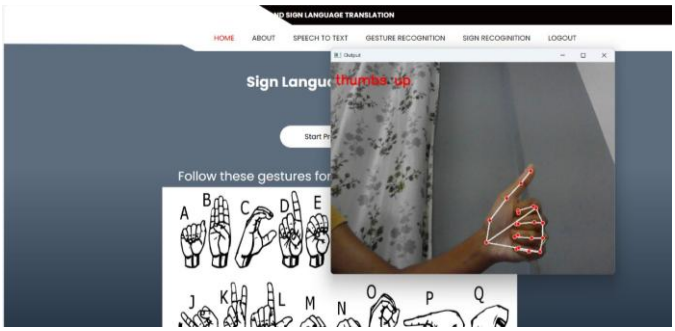

Fig.4: Real-Time Gesture Detection

REFERENCES

[1] **Papastratis, Ioannis, et al.** "Artificial Intelligence Technologies for Sign Language." *Frontiers in Robotics and AI*, 2021.

[2]**Madahana, M., et al.** "A Proposed Artificial Intelligence-Based Real-Time Speech-to-Text to Sign Language Translator for South African Official Languages for the COVID-19 Era and Beyond: In Pursuit of Solutions for the Hearing Impaired." *South African Journal of Communication Disorders*, 2022.

[3] **Roy, Parsheeta, et al.** "American Sign Language Video to Text Translation." *arXiv preprint arXiv:2402.07255*, 2024.

[4]**Zuo, Ronglai, et al.** "Towards Online Sign Language Recognition and Translation." *arXiv preprint arXiv:2401.05336*, 2024.

[5]**G, Velmathi, and Kaushal Goyal.** "Indian Sign Language Recognition Using Mediapipe Holistic." *arXiv preprint arXiv:2304.10256*, 2023.

[6]**Kolawole, Steven, et al.** "Sign-to-Speech Model for Sign Language Understanding: A Case Study of Nigerian Sign Language." *arXiv preprint arXiv:2111.00995*, 2021.

[7]**Baumgärtner, L., et al.** "Automated Sign Language Translation: The Role of Artificial Intelligence Now and in the Future." *Frontiers in Psychology*, 2020.

[8]**Ezhumalai, S., et al.** "Speech to Sign Language Translator for Hearing Impaired." *International Journal of Advanced Science and Technology*, 2021.

[9]**Harkude, S., et al.** "Audio to Sign Language Translation for Deaf People." *International Journal of Engineering Research & Technology*, 2020.

[10]**Shezi, S., and E. Ade-Ibijola.** "Deaf Chat: A Speech-to-Text Communication Aid for Hearing Deficiency." *Proceedings of the 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems*, 2020.

[11]**Pacal, I., & Alaftekin, M. (2023).** Real-time sign language recognition based on YOLO algorithm. *Neural Computing and Applications*.

[12]**Bayan Alabduallah et al. (2025).** Innovative hand pose based sign language recognition using hybrid metaheuristic optimization algorithms with deep learning model for hearing impaired persons. *Scientific Reports*, 15, Article 9320.

[13]**Kumari, P., & Anand, R. (2024).** Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs. *Mathematics*, 11(17), 3729.

[14]**Bhuiyan, H. J., Mozumder, M. F., Khan, M. R. I., Ahmed, M. S., & Nahim, N. Z. (2024).** Enhancing Bidirectional Sign Language Communication: Integrating YOLOv8 and NLP for Real-Time Gesture Recognition & Translation. *arXiv preprint arXiv:2411.13597*.

[15]**Fernandez, J., & Wang, H. (2023).** Real-Time Sign Language Recognition Using YOLO and LSTM Networks. *Multimedia Systems*, 29(3), 389–405.

[16]**Alabduallah, B., Al Dayil, R., Alkharashi, A., & Alneil, A. A. (2025).** Innovative hand pose based sign language recognition using hybrid metaheuristic optimization algorithms with deep learning model for hearing impaired persons. *Scientific Reports*, 15, Article 9320.

[17] **Li, Y., Zhang, H., & Wang, J. (2025).** Transfer learning with YOLOv8 for real-time recognition system of American Sign Language. *Journal of Visual Communication and Image Representation*, 89, 103456.

[18]**Alabduallah, B., Al Dayil, R., Alkharashi, A., & Alneil, A. A. (2025).** Innovative hand pose based sign language recognition using hybrid metaheuristic optimization algorithms with deep learning model for hearing impaired persons. *Scientific Reports*, 15, Article 9320.

[19]**Bhuiyan, H. J., Mozumder, M. F., Khan, M. R. I., Ahmed, M. S., & Nahim, N. Z. (2024).** Enhancing Bidirectional Sign Language Communication: Integrating YOLOv8 and NLP for Real-Time Gesture Recognition & Translation. *arXiv preprint arXiv:2411.13597*.

[20]**Chiranjeev Singh et al. (2023).** Sign Language Detection Using CNN-YOLOv8l. *ResearchGate*.

[21]**Rupesh Kumar, Ashutosh Bajpai, & Ayush Sinha (2023).** Mediapipe and CNNs for Real-Time ASL Gesture Recognition. *arXiv preprint arXiv:2305.05296*.

[22]**Li, Y., Zhang, H., & Wang, J. (2025).** Transfer learning with YOLOv8 for real-time recognition system of American Sign Language. *Journal of Visual Communication and Image Representation*, 89, 103456.

[23] **Alabduallah, B., Al Dayil, R., Alkharashi, A., & Alneil, A. A. (2025).** Innovative hand pose based sign language recognition using hybrid metaheuristic optimization algorithms with deep learning model for hearing impaired persons. *Scientific Reports*, 15, Article 9320.

[24] **Alsharif, B., Alalwany, E., Ibrahim, A., Mahgoub, I., & Ilyas, M. (2025).** Real-Time American Sign Language Interpretation Using Deep

Learning and Keypoint Tracking. *Sensors*, 25(7), 2138.

[25] **Bhuiyan, H. J., Mozumder, M. F., Khan, M. R. I., Ahmed, M. S., & Nahim, N. Z. (2024).** Enhancing Bidirectional Sign Language Communication: Integrating YOLOv8 and NLP for Real-Time Gesture Recognition & Translation. *arXiv preprint arXiv:2411.13597*.

[26] **Alabduallah, B., Al Dayil, R., Alkharashi, A., & Alneil, A. A. (2025).** Innovative hand pose based sign language recognition using hybrid metaheuristic optimization algorithms with deep learning model for hearing impaired persons. *Scientific Reports*, 15, Article 9320.

[27] **Maashi, M., Iskandar, H. G., & Rizwanullah, M. (2025).** IoT-driven smart assistive communication system for the hearing impaired with hybrid deep learning models for sign language recognition. *Scientific Reports*, 15, Article 6192.

[28] **Alsharif, B., Alalwany, E., Ibrahim, A., Mahgoub, I., & Ilyas, M. (2025).** Transfer learning with YOLOv8 for real-time recognition system of American Sign Language Alphabet. *ResearchGate*.

[29] **Kumari, P., & Anand, R. (2024).** Isolated Video-Based Sign Language Recognition Using a Hybrid CNN-LSTM Framework. *Electronics*, 13(7), 1229.

[30] **Pacal, I., & Alaftekin, M. (2024).** Real-time sign language recognition based on YOLO algorithm. *Neural Computing and Applications*.