

Instant Sign Language to Audio and Text Translation

Shubhangi Tirpude¹, Mohammad Saify Sheikh², Himanshu Shrigiriwar³, Soham Bedi⁴ and Rugved Mhatre⁵

¹⁻⁵Department of Computer Science Shri Ramdeobaba College of Engineering and Management Nagpur, India
Email: tirpudes@rknc.edu sheikhmm@rknc.edu shrigiriwarhg@rknc.edu bediss@rknc.edu mhatreru@rknc.edu

Abstract—In a world where gestures speak volumes and silence carries meaning, we introduce a groundbreaking innovation: a blend of technology and compassion that translates sign language into written text, and then to audio in real-time. Welcome to a realm where hands shape words, and inclusivity finds its voice. This report sheds light on the journey of creating a transformative solution, set to bridge communication gaps and champion accessibility for everyone. Our system is skilled at recognizing 26 letters, operating instantly to capture and interpret hand gestures accurately. By using advanced machine learning and computer vision, it identifies characters and forms words on-the-spot, allowing for complete sentences to be constructed effortlessly. This report not only explains the technical details of our solution but also highlights its profound impact on society. By breaking communication barriers for those with hearing impairments, our system promotes inclusivity, ensures equal access to information, and empowers users to participate fully in various social and professional contexts. The development of this real-time sign to-text conversion system represents a significant step towards creating an inclusive digital space. Its smart design, smooth operation, and societal benefits make it a symbol of progress, ready to redefine communication norms and enhance accessibility for all.

Index Terms— Gesture Recognition, Text Conversion, Machine Learning, Convolutional Neural Network.

I. INTRODUCTION

Sign language is a rich and expressive form of communication used by the deaf and hard of hearing community worldwide. However, the accessibility of sign language can be limited due to the lack of widespread understanding and interpretation. Sign language to text conversion with audio output is a groundbreaking technology that aims to bridge this communication gap by providing a means to translate sign language gestures into written text and spoken words in real-time.

Creating an advanced, real-time sign language to text conversion system represents a significant technological challenge with profound societal implications. This ambitious project seeks to address the pressing need for effective communication solutions for individuals with hearing impairments, aiming to bridge the gap between sign language users and those who rely on written or spoken language. By developing a robust and innovative solution, this endeavor aims to empower individuals with hearing impairments to communicate effortlessly and participate fully in various aspects of society, from education and employment to social interactions and beyond. The overarching goal is to foster inclusivity, promote accessibility, and enhance the quality of life for individuals with hearing impairments by enabling seamless and accurate translation of sign language gestures into written text in real-time.

The motivation behind real-time sign language to text conversion lies in addressing the Significant challenges and limitations present in current methods of interpreting and transcribing sign language. Individuals who rely on sign language communication face barriers in effectively expressing themselves and understanding others in various contexts, including education, healthcare, and public services. Traditional methods of transcription may not always capture the intricacies of sign language, leading to misinterpretations and misunderstandings. By leveraging advanced technologies such as machine learning algorithms and computer vision techniques, we aim to establish a more reliable and efficient system for real-time sign language to text conversion.

II. LITERATURE REVIEW

Researchers have proposed various methodologies for sign language recognition (SLR) systems, leveraging advancements in deep learning and computer vision techniques. Thakur et al. introduced a deep learning model trained on American Sign Language (ASL) data, achieving a notable accuracy improvement through Transfer Learning, reaching 98.7%. Kemkar et al. devised a real-time ASL recognition method utilizing Convolutional Neural Networks (CNNs), achieving accuracies of 99% and 96% on MNIST and ASL datasets respectively. Dharshini et al. suggested a comprehensive approach for recognizing various sign language poses, integrating Media Pipe Holistic, OpenCV, LSTM Neural Network, and translation services, achieving 98% accuracy. Abraham et al. developed a wearable device using flex sensors and an Android application for gesture recognition, enhancing communication accessibility. Adeyanju et al. conducted a literature review highlighting the importance of intelligent solutions in SLR, emphasizing the need for further research. Bagane et al. presented a technology for ASL recognition and translation into text using computer vision and machine learning algorithms. Akano et al. implemented image segmentation and feature detection techniques, achieving 78% accuracy through unsupervised learning. Zahid et al. contributed by creating a dataset for Urdu sign language (USL) and applying machine learning classifiers, achieving accuracies of 88%, 90%, and 84% using Random Forest, SVM, and K-NN respectively.

III. METHODOLOGY

The proposed methodology begins with the collection of a diverse dataset comprising various sign language gestures, covering alphabets and common phrases. Following data collection, preprocessing steps are employed to prepare the dataset for model training. This involves resizing and normalizing the images to enhance the model's ability to generalize. Subsequently, Convolutional Neural Networks (CNNs) are employed for model training, leveraging their capability to effectively recognize and interpret sign language gestures. Once the model is trained, focus shifts to the development of a user-friendly Graphical User Interface (GUI). The GUI facilitates seamless interaction with the trained model, enabling real-time sign language interpretation. Finally, the system undergoes rigorous testing to validate its performance before deployment, ensuring that it effectively promotes inclusivity and accessibility for users.

A. Data Collection

At the outset of the project, a comprehensive collection of diverse hand gesture images was prioritized to serve as foundational data for the sign language recognition system. A specialized Python program was employed to systematically acquire images representing the complete English alphabet and common phrases.

1) Dataset Creation: Extensive efforts were dedicated to meticulously gather a diverse array of hand gesture images, ensuring comprehensive coverage across various gestures. Utilizing the Python program facilitated efficient image capture, encompassing a wide spectrum of hand positions and angles, thereby enriching the dataset and enhancing model training.

2) Preprocessing Techniques: Cutting-edge computer vision methodologies were employed to preprocess the collected images, aiming to improve their quality and standardize characteristics. Techniques such as grayscale conversion and resizing were strategically applied to promote uniformity and mitigate variability, laying the foundation for consistent model performance.

3) Dataset Partitioning: Recognizing the significance of model evaluation, the dataset was judiciously partitioned into training and testing subsets. 80% of the images were allocated for model training, while the remaining 20% constituted the test set. This meticulous partitioning ensured thorough model evaluation and robustness across diverse scenarios.

B. Modeling Approach

Using our carefully collected dataset, we built a special computer program called a Convolutional Neural Network (CNN) for recognizing sign language. We used a powerful tool called Keras to design this program, making sure it could understand even the most complex hand movements and match them with the right letters and numbers accurately and quickly.

1) Convolutional Layers: The core of our CNN model comprises multiple convolutional layers meticulously crafted to extract and capture spatial features from input hand gesture images. Convolutions, fundamental to CNNs, execute localized operations to detect patterns and features within the input data. Mathematically, the convolution operation is expressed as

$$\text{Output} = \text{Convolution}(\text{Input}, \text{Kernel}) + \text{Bias}.$$

Here, the kernel represents a learnable parameter matrix applied to the input data, and the bias term introduces a translation invariant component to the output. By cascading multiple convolutional layers, our model hierarchically extracts increasingly abstract and complex features, facilitating robust pattern recognition and classification.

2) Max-Pooling and Dropout Layers: Following each convolutional layer, we integrated max-pooling and dropout layers to enhance feature extraction and alleviate overfitting. Max pooling layers down sample feature maps, retaining salient information while discarding redundant details, thus reducing computational complexity and increasing translational invariance. Additionally, strategically inserted dropout layers regularize the model by randomly deactivating a fraction of neurons during training, encouraging the learning of more robust and generalizable features. This selective dropout encourages the model to avoid reliance on specific neurons, enhancing its ability to generalize to unseen data.

3) Fully Connected Layers: Towards the network's end, fully connected layers were incorporated to facilitate high level abstraction and pattern recognition. These dense layers integrate features extracted from convolutional layers, enabling the model to capture complex relationships and correlations within the data. Mathematically, the output of a fully connected layer is represented as

$$\text{Output} = \text{Activation}(\text{Weights} \times \text{Input} + \text{Bias}).$$

Here, weights denote learnable parameters optimized during training to minimize the model's loss function. By applying non-linear activation functions such as ReLU or sigmoid, fully connected layers introduce non-linearity, enabling the model to learn complex decision boundaries and make accurate predictions. Fully connected layer is demonstrated in Fig 1.

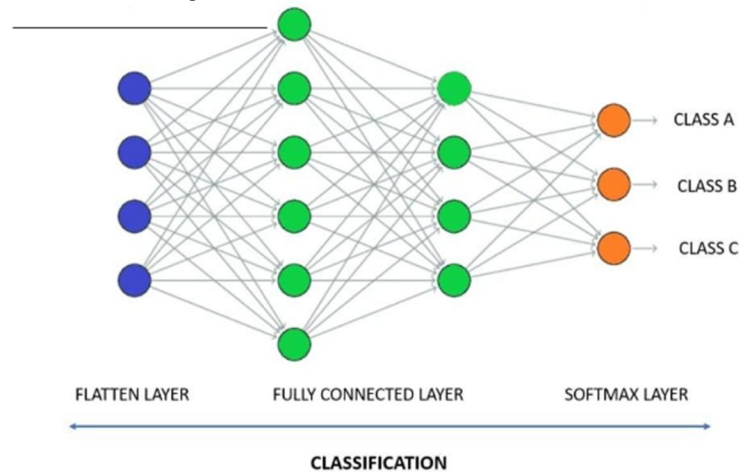


Fig. 1. Fully Connected Layer

4) Output Layer: At the culmination of our CNN architecture lies the output layer, employing the softmax activation function to assign probabilities to each class, facilitating final classification decisions. This function ensures a probability distribution across multiple classes, with a sum of probabilities equal to one. Our meticulously crafted CNN architecture, integrating convolutional, max-pooling, dropout, and fully connected layers, represents a sophisticated framework for sign language recognition, poised to revolutionize real-time communication for individuals with diverse needs.

C. Real-Time Inference System

At the core of our project is a real-time inference system, comprising a user-friendly interface for converting sign language into text. This intuitive application utilizes Python's tkinter library for GUI construction and OpenCV for video processing. Upon initiation, users are presented with a live video feed capturing hand gestures, featuring:

- **Letter Row:** Displays the currently recognized letter based on hand gestures.
- **Word Row:** Shows the ongoing formation of words as gestures are made.
- **Sentence Row:** Displays the evolving sentence constructed from successive gestures.
- **Audio Button:** Converts interpreted text into speech for enhanced accessibility.

The system rapidly analyzes hand movements, identifies characters, and updates displayed letters, words, and sentences in real-time, facilitating natural and fluid communication while promoting inclusivity and accessibility.

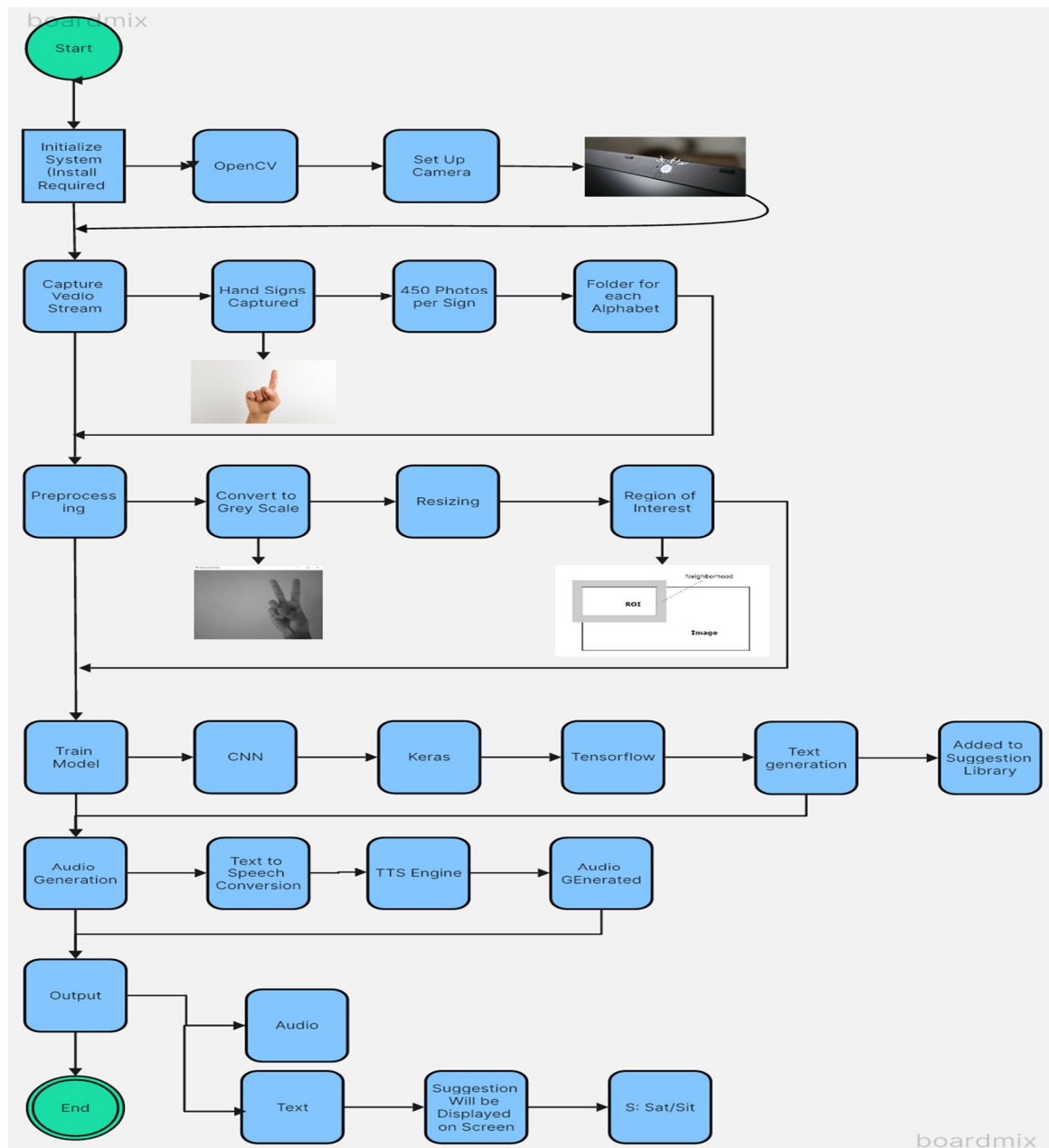


Fig. 2. Proposed Working Flowchart

IV. IMPLEMENTATION

The implementation phase translates the conceptual design into a practical solution, detailing steps to execute the plan, Fig. 2. Proposed Working Flowchart configure infrastructure, integrate software components, and conduct rigorous testing.

- 1) Data Collection and Preprocessing: A dataset of sign language gestures was collected using a Python script named collect.py, leveraging OpenCV for image capture. Each labeled image underwent preprocessing techniques including resizing, normalization, and augmentation to enhance dataset quality and diversity.
- 2) Model Training: The collected dataset trained a CNN model for gesture recognition using TensorFlow with Keras. Hyperparameters were fine-tuned, and the dataset was split into training and validation set.
- 3) Real-time Sign Language Detection: A deep learning model was trained on a dataset of sign language images for real-time detection. Using OpenCV, the webcam captured video, processed frames, and the trained model predicted gestures, displayed via a Tkinter GUI interface.

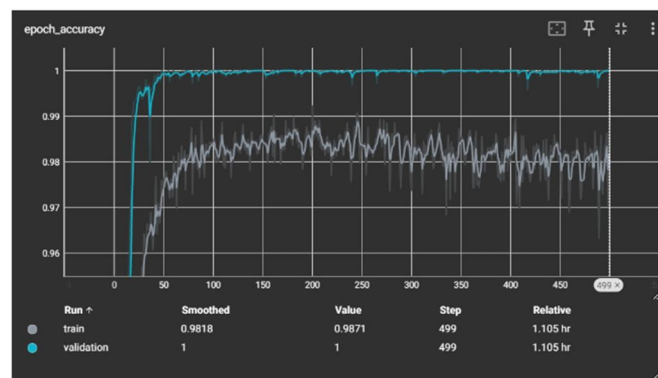


Fig. 3. Accuracy vs Epoch

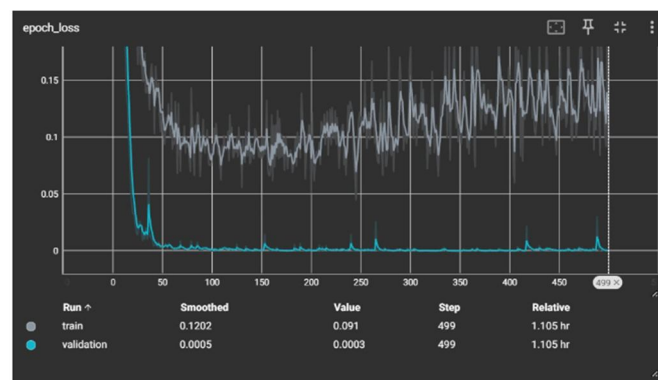


Fig. 4. Loss vs epoch

- 4) Gesture Recognition and Text Prediction: The trained CNN model recognized gestures in real-time, predicting corresponding letters appended to a queue for word and sentence formation, with a blank symbol introduced for readability.
- 5) Suggestion Feature: A suggestion feature leveraging MongoDB stored a letter list, dynamically querying to suggest words or phrases, enhancing communication efficiency and user experience.
- 6) Graphical User Interface (GUI) Implementation: A Tkinter GUI displayed real-time video feed, detected hand regions, recognized letters, and formed words/sentences, enabling user interaction and observation of sign language gesture conversion into text.

V. RESULTS

This section provides a comprehensive analysis of the performance and efficacy of our real-time sign-to-text conversion system. Leveraging a meticulously crafted Convolutional Neural Network (CNN) architecture and bespoke graphical user interface (GUI), our system embodies cutting-edge technology and innovative design.

Through experiments and evaluations, we delve into quantitative and qualitative aspects, shedding light on robustness, accuracy, and real-world applicability.

1) Training Dynamics: The training dynamics of our CNN model were analyzed using TensorBoard, offering insights into key parameters such as loss, accuracy, and learning rate. Dynamic graphs and visualizations elucidated the model's convergence behavior and optimization trajectory.

2) Model Performance: Rigorous evaluation of our real-time sign language recognition system was conducted through extensive testing on diverse datasets. By subjecting the system to real-world conditions, valuable insights into robustness, accuracy, and adaptability were elicited, affirming its efficacy as a reliable communication aid.

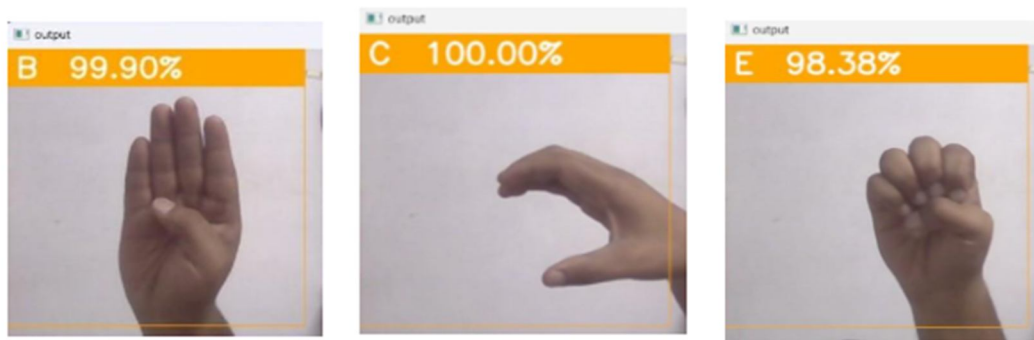


Fig. 5. Results of alphabet Recognition

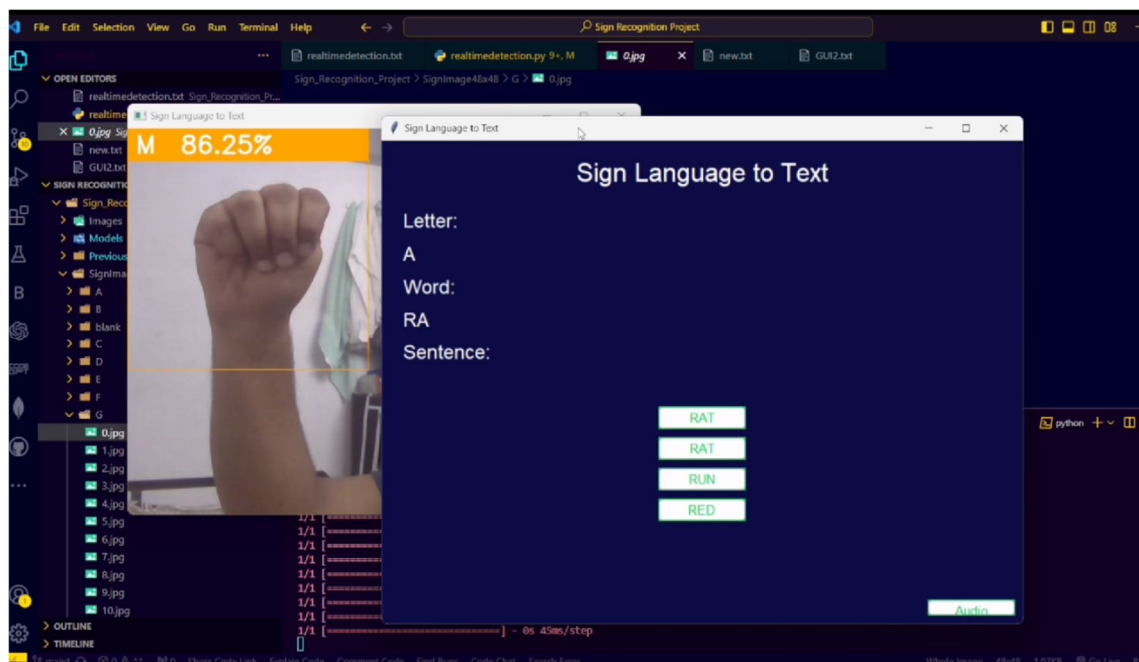


Fig. 6. GUI Output

VI. CONCLUSION AND FUTURE SCOPE

Our sign language to text conversion model utilizing CNN technology has demonstrated promising results in accurately translating sign language gestures into text. Through extensive experimentation and evaluation, we have shown its effectiveness in bridging the communication gap between individuals with hearing impairments and those not proficient in sign language. Leveraging convolutional neural networks, we achieved robustness and efficiency in recognizing complex hand gestures and translating them into meaningful text and sound representations.

Despite the advancements made, there are avenues for future research and development to further enhance capabilities and applicability:

A. Improving Accuracy

Continuously refining the model architecture and training methodologies can lead to higher accuracy in gesture recognition and translation. Incorporating larger and more diverse datasets can capture a wider range of sign language variations and nuances.

B. Multimodal Integration

Exploring the integration of additional modalities such as facial expressions and body movements alongside hand gestures can enrich context and accuracy. This holistic approach can better capture nuances and emotions conveyed through sign language.

REFERENCES

- [1] S. Thakar, S. Shah, B. Shah and A. V. Nimkar, "Sign Language to Text Conversion in Real Time using Transfer Learning," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-5, doi: 10.1109/GCAT55367.2022.9971953.
- [2] T. Kemkar, V. Rai and B. Verma, "Sign Language to Text Conversion using Hand Gesture Recognition," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 1580-1587, doi: 10.1109/ICCES57224.2023.10192820.
- [3] P. Bagane, M. Thawani, P. Singh, R. Ahmad, R. Mital, O. Amenu, "Breaking the Silence: An innovative ASL-to-Text Conversion System Leveraging Computer Vision and Machine Learning for Enhanced Communication," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, pp. 246-255, 2024.
- [4] A. Abraham, V. Rohini, "Real-time conversion of sign language to speech and prediction of gestures using Artificial Neural Network," Procedia Computer Science, vol. 143, pp. 587-594, 2018, doi: 10.1016/j.procs.2018.10.435.
- [5] Adeyanju, O.O. Bello, M.A. Adegboye, "Machine learning methods for sign language recognition: A critical review and analysis," Intelligent Systems with Applications, vol. 12, 2021, Article 200056, ISSN 2667- 3053.
- [6] V. Akano, A. Olamiti, "Conversion of Sign Language To Text And Speech Using Machine Learning Techniques," JOURNAL OF RESEARCH AND REVIEW IN SCIENCE, vol. 5, pp. 58-65, 2018.
- [7] A.M. Buttar, U. Ahmad, A.H. Gumaei, A. Assiri, M.A. Akbar, B.F. Alkhamees, "Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs," Mathematics, vol. 11, no. 16, article 3729, 2023.
- [8] H. Zahid, M. Rashid, S.A. Syed, R. Ullah, M. Asif, M. Khan, A. Abdul Mujeeb, A. Haider Khan, "A computer vision-based system for recognition and classification of Urdu sign language dataset," PeerJ Computer Sci, vol. 8, article e1174, 2022, doi: 10.7717/peerj-cs.1174.
- [9] H. Kumar, M.K. Sharma, R. Kumar, K.S. Bisht, A. Kumar, R. Jain, P. Nagrath, P. Singh, "Sign language detection and conversion to text using CNN and OpenCV," AIP Conf. Proc., vol. 2555, article 040016, 2022.
- [10] A. Deshpande, A. Shriwas, V. Deshmukh, S. Kale, "Sign Language Recognition System using CNN," 2023 International Conference on Intelligent And Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, 2023, pp. 906-911, doi: 10.1109/IITCEE57236.2023.10091051.