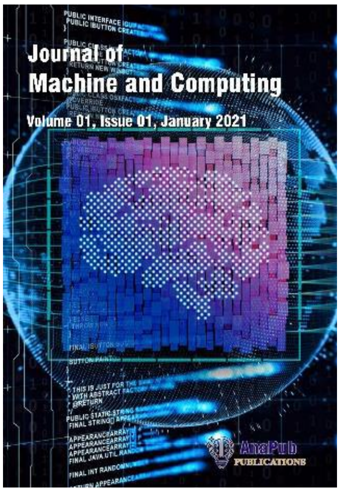
Journal Pre-proof

Develop an Ensemble Transfer Learning with Hybrid Vision Transformers with Convolutions for Enhancing Indian Sign Language Recognition



Suresh Anand M, Mong-Fong Horng and Chin-Shiuh Shieh

DOI: 10.53759/7669/jmc202505185

Reference: JMC202505185

Journal: Journal of Machine and Computing.

Received 02 May 2025

Revised from 16 July 2025

Accepted 03 August 2025

Please cite this article as: Suresh Anand M, Mong-Fong Horng and Chin-Shiuh Shieh, "Develop an Ensemble Transfer Learning with Hybrid Vision Transformers with Convolutions for Enhancing Indian Sign Language Recognition", Journal of Machine and Computing. (2025). Doi: https://doi.org/10.53759/7669/jmc202505185.

This PDF file contains an article that has undergone certain improvements after acceptance. These enhancements include the addition of a cover page, metadata, and formatting changes aimed at enhancing readability. However, it is important to note that this version is not considered the final authoritative version of the article.

Prior to its official publication, this version will undergo further stages of refinement, such as copyediting, typesetting, and comprehensive review. These processes are implemented to ensure the article's final form is of the highest quality. The purpose of sharing this version is to offer early visibility of the article's content to readers.

Please be aware that throughout the production process, it is possible that errors or discrepancies may be identified, which could impact the content. Additionally, all legal disclaimers applicable to the journal remain in effect.

AnaPub
PUBLICATIONS

Develop an Ensemble Transfer Learning with Hybrid Vision Transformers with Convolutions for Enhancing Indian Sign Language Recognition

[1]Suresh Anand.M, [2] Mong-Fong Horng, [3] Chin-Shiuh Shieh

[1]Assistant Professor, Department of Computing Technologies, School of Computing, SRM Institute of Science & Technology,Kattankulathur, India -603203.

[2]Professor, Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Taiwan

[3]Professor, Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Taiwan

suresh.anandm@gmail.com, mfhorng@nkust.edu.tw, csshieh@nkust.edu.tw

Abstract

Indian Sign Language (ISL) identification methods play a central role in enhancing communication between hearing-impaired and non-impaired individuals within their community. However, modern ISL identification algorithms face challenges due to hand gesture variability, complex visual settings, and limited official annotations. This study proposes a Hybrid Vision Transformer with Convolutions (HVTC) combined with Ensemble Transfer Learning (ETL), incorporating advanced transfer learning methods such as Adaptive Lightweight DenseNet, VGG19, and XceptionNet for Multi-Task Learning, along with ResNet with Dynamic Depth and MobileNetV3 with Attention Mechanisms to improve ISL recognition accuracy. Four primary challenges affect ISL recognition: obstructions in the camera view, inconsistent lighting conditions, visually similar motions that are difficult to distinguish, and the need for extensive labeled datasets for deep learning systems. The ETL-HVTC processing method effectively extracts spatial-temporal motion data by leveraging sophisticated neural network algorithms. Transfer learning reduces dependency on large datasets, while the ensemble approach integrates multiple predictive models to enhance model stability. A robust ISL recognition algorithm should prioritize real-time capabilities, high recognition accuracy, and an expanded application scope. Secure gesture dataset pre-processing enables the optimization of hybrid ViT Large Model-CNN models, where collaborative learning ensures reliable classification outcomes. Experimental results demonstrate that the proposed ETL-HVTC system outperforms independent ViT Large Model and existing CNN models on ISL benchmark databases in terms of precision, recall, F1-score, and accuracy. The implementation approach yields fast and effective results, facilitating the

development of assistive devices that promote more inclusive communication for individuals with hearing impairments.

Keywords

Indian Sign Language Recognition, Vision Transformers, Convolutional Neural Networks, Transfer Learning, Ensemble Learning, Deep Learning, Hybrid Models, Gesture Recognition, Assistive Communication, Multimodal Feature Extraction.

1. Introduction

Translation to sign phrases and translating from sign phrases are two distinct categories in the field of sign language translation. This classification highlights the various methods and tools that facilitate interactions between the general public and individuals with hearing impairments. The deaf community communicates through sign language using visual cues and motions, including manual signals, body movements, and facial expressions [1]. For communication, speech-impaired and visually impaired individuals use Indian Sign Language (ISL), a motion-based form of speech. This highly complex communication system relies on distinct hand gestures, communication styles, and situational responses. ISL differs from all other spoken languages in India in terms of vocabulary and grammatical structure [2]. In today's rapidly evolving technological landscape, ensuring accessibility for everyone, including individuals with hearing loss, remains a top priority. Sign Language Recognition (SLR) enhances communication between the hearing-impaired population and the general public [3].

Existing identification systems have long struggled due to the dynamic and complex nature of signing motions. Previous studies have primarily focused on developing sensor-based algorithms and static rule-based techniques lack flexibility in accommodating different signing techniques and variations among individuals. The identification of sign language has significantly advanced with recent developments in Machine Learning (ML) And Deep Learning (DL) through Artificial Intelligence (AI) and Convolutional Neural Networks (CNNs) [4]. Tools that detect and analyse hand positions, facial expressions, and body signals now operate more effectively due to vision-based recognition methods that integrate image processing with pattern recognition techniques. Many existing technologies still face challenges related to precision, user-friendliness, and real-time adaptability. While SLR systems show promise for improvement often fail to accommodate various signing approaches, linguistic variations, and individual preferences [5]. Most contemporary SLR systems support only a limited set of sign dialects and languages, making universal interpretation across different users difficult. Another unresolved challenge is achieving real-time gesture recognition without compromising accuracy. User-friendly interface is essential to ensure accessibility for individuals with varying levels of technical expertise, even though many existing systems tend to favour technical complexity [6].

The process of converting spoken words or text into sign language falls into two main categories. The first, text-to-sign language conversion, aims to translate written content into corresponding signs, enabling individuals who are deaf or hard of hearing to access textual information. This process typically involves Natural Language Processing (NLP) techniques to interpret text data and generate appropriate sign language visualizations [7]. On the other hand, speech-to-sign language interpretation focuses on translating spoken language into sign language, ensuring effective communication between individuals who rely on signing and those who communicate verbally [8].

Sophisticated voice recognition techniques are employed in this continuously evolving process to capture spoken words and convert them into corresponding signals. The goal is to provide real-time translation, enabling seamless interaction between individuals who communicate through speech and those who use sign language. The second category discussed in this paper focuses on identification systems based on vision and sensors, which facilitate sign language translation into other languages [9]. Vision-based identification employs computer vision techniques to analyse and interpret sign movements captured by cameras or other visual input devices. This approach enhances real-time sign recognition, allowing individuals with hearing impairments to communicate more efficiently and effectively [10]. Sensor-based identification, on the other hand, extends the interpretation of sign language by capturing various aspects of sign communication shown in Figures 1 and 2. This method utilizes sensors to track body positions, hand movements, and facial expressions, providing a comprehensive understanding of sign gestures. By integrating sensor data, translation accuracy and complexity are significantly improved, effectively addressing the nuances of sign language [11].

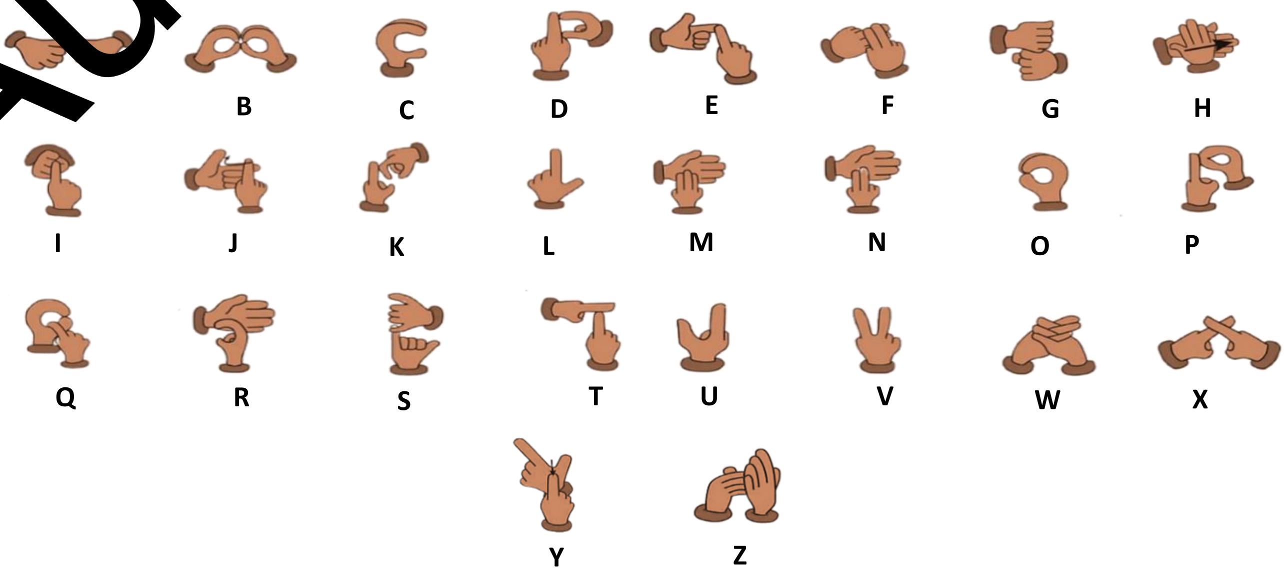


Figure 1. ISL alphabets

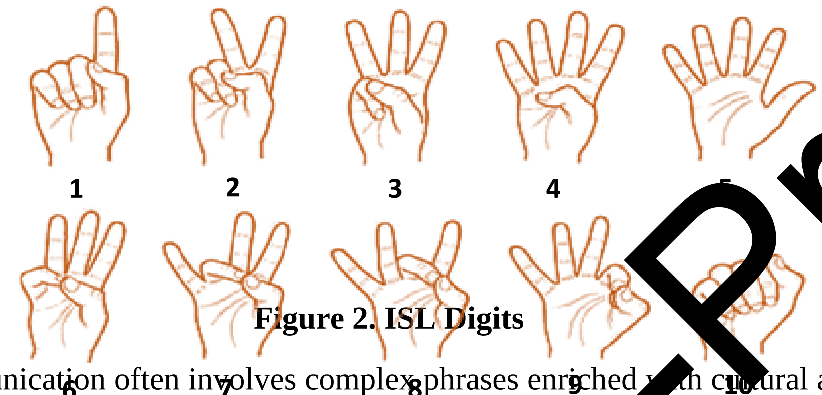


Figure 2. ISL Digits

Informal communication often involves complex phrases enriched with cultural and linguistic nuances. Existing sign language comprehension and translation methods rely heavily on prebuilt algorithms and predefined datasets. These approaches struggle to adapt to the dynamic nature of sign languages. In this study, Random Forest Classification algorithms were employed to effectively recognize gestures, while large language models were utilized to enhance context-aware translations [12]. This study introduces a text-based intermediary representation that bridges the gap between movement generation and detection. This translational intermediary not only ensures more accurate rendering but also allows for adaptable interpretation while preserving the original expression's meaning and cultural intricacies [13].

One of the key contributions of this model is its ability to mitigate linguistic discrepancies between American Sign Language (ASL) and ISL differ significantly in word order, grammatical structures, and situational expressions. By addressing these variations, the framework enhances the fidelity of cross-language sign translation. Employing RIFE-Net to generate ISL movements from written translations results in fluid and naturalistic motion displays [14]. RIFE-Net not only accurately reproduces ISL gestures demonstrates exceptional capability in handling variations in movement sequencing. The program's architecture integrates advanced recognition, translation, and synthesis components, positions it at the forefront of sign language translation technology. By combining real-time gesture recognition with culturally aware processing and adaptive movement synthesis, this framework establishes a new standard in sign language translation systems [15].

Vision-based SLR is more prevalent than sensor-based SLR due to its real-time applicability. Deep neural networks and LSTM-CNN have been widely utilized in SLR development. SLR remains constrained by the complexity of sign language, environmental conditions, and dataset integrity. The researchers emphasized the need to develop robust universal frameworks capable of handling diverse signers and varying contexts [16]. After reviewing SLR methodologies, proposed more reliable techniques that could operate in different environments, accommodate larger vocabularies, and address key challenges that hinder SLR's practical implementation. Concluded that future studies should integrate multiple modalities to improve reliability. Lack of extensive datasets collected through continuous sign language recordings or smartphone cameras limits existing advancements in the field. The available data for SLR can be categorized into two main types: time-series analysis data and static information [17]. Classification techniques are generally classified into modern deep learning approaches, such as LSTM, CNN and existing machine learning methods such as Support Vector Machines (SVM) and Hidden Markov Models (HMM). Input hardware can be broadly classified into recording devices and sensor-based detectors. These systems can accurately recognize and interpret sign movements using machine learning techniques [18]. Automated vision technology has further enhanced accessibility for individuals with hearing difficulties by enabling the development of sign language translation devices capable of converting spoken speech into sign language and vice versa. The widespread adoption of vision technology plays a crucial role in making communication more inclusive and user-friendly [19].

The successful implementation of AI-powered sign language recognition relies on effective feature extraction methods, which form the foundation of machine learning techniques. Training AI models involves extracting meaningful information from sign movement data, typically obtained from images and recorded videos. The detection process utilizes key methods such as Histogram of Oriented Gradients (HOG), CNN, and Scale-Invariant Feature Transform (SIFT), along with other commonly used techniques for feature extraction. These methods aim to improve the accuracy of sign language identification, thereby enhancing accessibility options for individuals with hearing impairments [20].

1.1 Problem Statement

The detection of ISL encounters multiple hurdles caused by complicated hand expressions and speed variations together with physical barriers and individual signature variations. Standard artificial intelligence together with deep learning present recognition issues because they fail to identify temporal and spatial relationships properly. The development of reliable ISL identification systems faces two main obstacles from insufficient big designated data sets and the need for instant processing. The solution proposed to these issues incorporates highly developed Learning strategies merge ViT Large Model-CNN and ensemble transferred learning techniques. A combination approach enables the system to process ISL movements with intricate differences effectively thus assuring effective visual language detection.

1.2 Motivation

The research examines ISL recognition because the deaf community needs accurate sign language translation to bridge their communication gap with the general population. Existing recognition systems struggle with practical usefulness because they perform poorly under conditions of hand closures and complicated hand motions while allowing diverse methods of communication. The fast advancements in neural networks particularly ViT Large Model-CNN create opportunities to achieve better accuracy and quicker processing in ISL identification systems. The research aims to build an adaptable ISL movement identifier by using collection transferred learning within a hybrid structure of ViT Large Model-CNN. Such integration will support effective communication between people with hearing disabilities by fostering inclusive social interaction.

Key contribution of the paper are as follows:

- To Combines Vision Transformer and Convolutions to enhance spatial-temporal feature extraction for improved Indian Sign Language recognition accuracy.
- Utilizes Adaptive DenseNet, VGG-19, XceptionNet, ResNet with Dynamic Depth, and MobileNetV3 for better generalization and performance.
- To overcome camera obstructions, lighting inconsistencies, gesture similarities, and dataset limitations through advanced deep learning techniques and transfer learning strategies.
- To implements optimized ViT Large Model-CNN hybridization with secure dataset pre-processing to support real-time ISL recognition and classification.
- To outperforms existing CNN and ViT models in precision, recall, F1-score, and accuracy for reliable, inclusive communication solutions.

2. Related Works

Sign languages play a crucial role in communication among deaf and mute individuals, and researchers have recently focused more on their identification and translation needs. This review examines how sign language identification techniques operate while addressing translation-related challenges. Developed a system using the Natural Language Toolkit framework to demonstrate how different linguistic groups produce sign language utterances, confirming the importance of linguistic analysis in sign language translation [21]. Comprehensive review on the challenges and advancements in deep learning-based sign language recognition. Developed an algorithm for sign language detection, providing a foundation for research advancements in this field. Highlighted the effectiveness of converters in interpreting signs by proposing a Transformer Network for video-to-text translation [22]. Proposed the design of real-time vernacular spoken language identification systems by integrating MediaPipe and AI showcasing immediate application prospects in this field. Research on sign language technologies for deaf communication continues with developing an advanced machine learning-based full-duplex sign language messaging system capable of handling multiple sign languages. Introduced the enhanced 3D-ResNet sign language identification method incorporating novel features to improve gesture recognition. Provided an extensive discussion on SLR challenges and potential solution approaches in their research [23].

A multi-headed CNN was implemented to develop a fusion method for SLR integrating hand and image landmarks to enhance gesture identification algorithms. A user-independent approach to ASL word recognition was presented using PCANet, operating in conjunction with the Microsoft Kinect. Researchers applied recursive neural networks to process GMU-ASL51 benchmarks, as outlined [24]. 26 ISL indicators were analysed using a Dynamic Time Warping (DTW) method achieving a [illegible]7.2% accuracy rate. Introduced a technique that integrates global and local ISL indicator data using the Axis of Least Inertia methodology. A 3D local characteristic integration method was also employed, relying on 3D key point analysis [24]. Using a multi-class machine learning approach achieved an 86.16% real-time recognition rate for 37 ISL indicators. Combination of DWT and HMM was utilized to identify 500 samples from 10 ISL phrases, resulting in a 91% reliability rate. Obtained 90% precision in recognizing 24 ISL hand motions using the DTW approach [25].

Emphasized that motion identification and feature extraction remain crucial in designing SLR systems. Developed an ISL translation model based on gesture recognition algorithms. Inspired by the exceptional translation capabilities of Large Language Models (LLMs), proposed leveraging commercially available LLMs to address complex Sign Language Translation tasks. Emphasized the importance of investigating gloss-free approaches, arguing that such methods could significantly reduce annotation time while promoting the development of more precise and comprehensive sign language translation frameworks [26]. Introduced the first large-scale multilingual Sign Language Processing (SLP) model, SIGNLLM. Developed using the Prompt2Sign database, SIGNLLM is capable of generating skeletal postures of sign language characters from text or prompts in eight distinct languages. Computerized translation technologies enable deaf or mute individuals to communicate effectively even without prior knowledge of sign language by converting gestures into spoken or written language [27]. Developing a computerized system that can translate between ISL and conventional languages is essential in today's world. Such a system is crucial for enhancing communication between the general public and individuals with hearing or speech impairments particularly when accessing essential services such as transportation, financial institutions, and ticketing systems [28].

To enhance human-computer interaction, propose a novel feature extraction and selection method for identifying ISL gestures. This method leverages advanced algorithms and seamlessly integrates structural characteristics. The proposed system employs only standard digital cameras, eliminating the need for specialized wearable devices. For optimal performance, each submitted image should exclusively depict a numerical sign, ensuring the system's ability to accurately translate these representations into text. To facilitate real-time ISL sign recognition, developed a comprehensive sign library consisting of 5,000 images, with 500 images dedicated to each of the nine numerical signs. In classifier evaluation, k-Nearest Neighbors (k-NN) demonstrated superior classification accuracy compared to Naïve Bayes [29].

The challenges posed by these additional features, combined with the regional variations in spoken languages, have resulted in limited research in the field of ISL. Effective communication with ISL users typically requires learning the language. While peer groups are the most common environment for learning sign language, there is a scarcity of instructional resources in this area. As a result, acquiring sign language proficiency is a significant challenge. The need for finger-spelling arises in the early stages of learning sign language, particularly when there is no equivalent sign for a word [30]. Existing SLR methods often rely on expensive third-party sensors. The data will then be integrated into supervised learning approaches with the validation set including images of different individuals from the training set. This methodology distinguishes our work from existing research shown in Figure 3 [30]. The primary objective of these systems is to enable seamless communication between these two modes. The foundational concept behind the system's introduction is an intelligent architecture capable of converting spoken languages such as English, into text and vice versa. Researchers emphasize how sign language translation technologies can aid the deaf community by improving communication, facilitating knowledge exchange, and creating better job opportunities [31]. The study addresses challenges in voice recognition, specifically focusing on the use of Mel-Frequency Cepstral Coefficients (MFCCs) to extract speech features. Key issues tackled by the proposed approach include the transition from speaker-dependent to speaker-independent speech recognition and the lack of comprehensive sound datasets for identification. The process phases include pre-processing, signal conditioning, feature extraction using Cepstral coefficients, and segmentation [32].

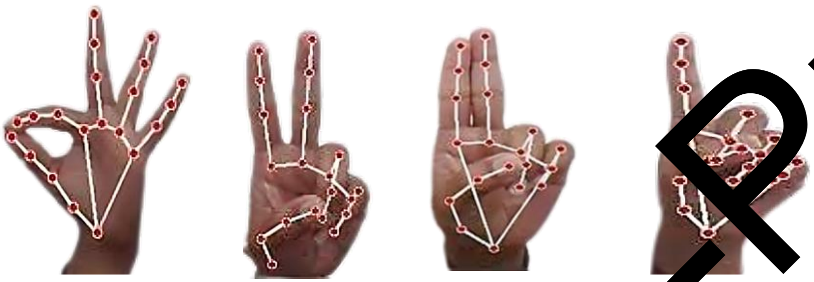


Figure 3: Recognition using Machine Learning

Systematic research successfully identified key elements within sign language-to-text translation structures, emphasizing the application of deep learning techniques. This approach proved highly effective in recognizing human gesture input and delivering precise translations.

As part of the study, refined an initial set of 40 relevant studies to 20 papers, specifically focusing on deep learning-based sign language translation. This selection was achieved through a two-step screening process. Among the methodologies analysed, CNN emerged as the dominant technique, accounting for 70% of the total study time. Connectionist Temporal Classification (CTC) followed with 20%, while Deep Belief Networks (DBN) contributed 10%. The findings of this paper provide valuable insights for researchers interested in leveraging deep learning techniques for sign language translation and identification.

Research Gap

Existing ISL recognition techniques have a number of drawbacks, although notable progress in the field. Many convolutional neural network systems rely on CNNs to extract spatial features because these networks prove valuable for spatial feature extraction. These networks struggle to understand relationships between data points within their contexts as well as to track dependent long-term hand motions. Both the design focus and limited adaptability characterize most ISL recognition models that target specific databases as they struggle to work effectively on diverse hand introductions and different signature styles in various contextual factors.

Generalized performance using transfer learning techniques is underutilized in ISL identification systems because it reduces the ability to learn from limited training datasets in novel databases.

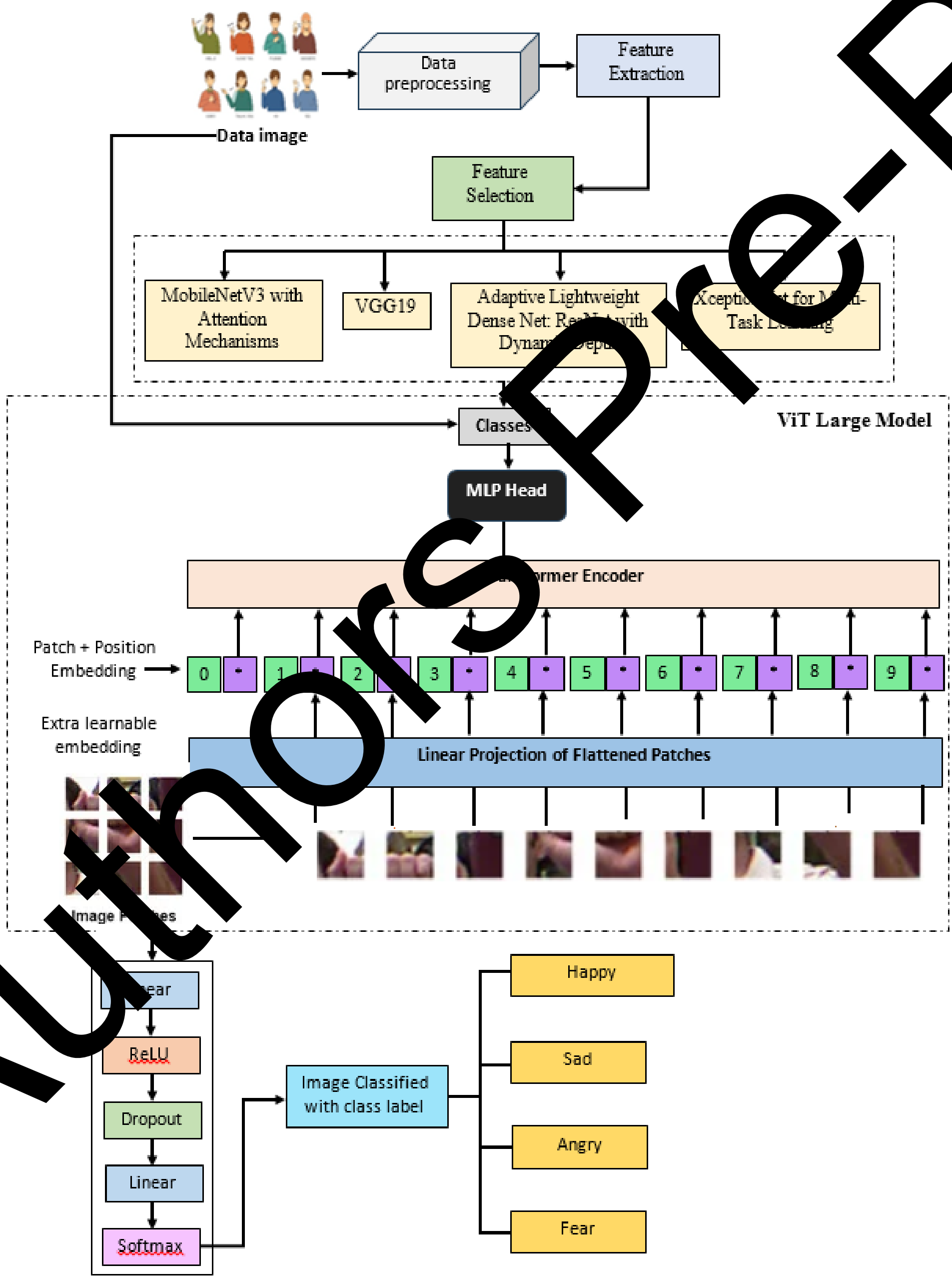


Figure 4: Proposed Architecture

3. Materials and Methods

The proposed ETL-HVTC framework with ViT Large model is designed to address the limitations of existing ISL recognition systems by combining the strengths of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) in a structured ensemble is shown in Figure 4. ViTs Large in capturing long-range dependencies and contextual relationships, making them ideal for understanding hand gestures in dynamic environments. In contrast, CNNs specialize in learning localized spatial features, ensuring robust feature extraction from complex sign language representations. The ensemble leverages Adaptive Lightweight DenseNet for efficient feature propagation, VGG19 for deep hierarchical feature extraction, and XceptionNet for depthwise separable convolutions that enhance computational efficiency. ResNet with Dynamic Depth enables adaptive learning by dynamically adjusting network depth based on input complexity, while MobileNetV3 with Attention Mechanisms enhances real-time recognition through lightweight yet powerful representations.

The Multi-Task Learning (MTL) strategy integrates gesture recognition, facial expression analysis, and environmental context awareness, significantly improving robustness against occlusions, lighting variations, and visually ambiguous gestures. Transfer learning ensures reduced dependency on large-scale labeled datasets while maintaining high recognition accuracy. Compared to existing CNN-based and ViT-based models, the proposed hybrid approach exploits the global contextual reasoning of ViTs and the precise local feature extraction of CNNs, leading to superior generalization.

3.1 Dataset Description

The ISL Movement Dataset supports ISL identification by incorporating hand motions, alphabet letters, numerals, phrases, actiViT Large Modelies, and emotional expressions, as summarized in Table 1. It consists of high-quality RGB images with varying resolutions 128×128, 256×256, and 512×512 pixels containing 50,000 to 200,000 samples. Each image includes posture key points, hand markers, and bounding boxes, enhancing identification accuracy. The dataset accommodates diverse signers with distinct hand profiles, skin tones, and physical attributes, ensuring robustness. This dataset enables the development of transfer learning ensemble models by integrating hybrid neural networks and convolutional architectures, facilitating precise and efficient ISL translation across various users and real-world scenarios.

Table 1: Dataset Description

Attribute	Description
Dataset Name	Indian Sign Language (ISL) Gesture Dataset
Source	Collected from real-time signers, public ISL datasets, and annotated video recordings
Number of Classes	50–200 (varies based on dataset used)
Categories	Alphabets, Numbers, Common Words, Emotions, Actions, Gestures
Total Samples	50,000 – 200,000 images/videos
Data Type	RGB Images and Video Frames
Resolution	128×128, 256×256, 512×512 pixels (varies)
Formats	JPG, PNG (for images), MP4, AVI (for videos)
Annotations	Bounding boxes, Hand landmarks, Pose key points
Pre-processing Steps	Image resizing, Normalization, Data Augmentation (rotation, flipping, brightness adjustment)
Splitting Ratio	Training (70%), Validation (15%), Testing (15%)
Challenges in Data	Variations in lighting, background noise, occlusions, signer-dependent variations
Augmentation Techniques	Random cropping, Gaussian noise, Adaptive histogram equalization, Motion blur simulation

Table 2 aids in modelling development, verification, and evaluation by representing a variety of spoken motions, groupings, and environmental situations. The dataset is helpful for deep learning-based language recognition algorithms since it includes annotation such posture important details and boxes with boundaries.

Table 2: Sample Data

Sample ID	Gesture	Category	Image Resolution	Signer ID	Pose Keypoints	Bounding Box (x, y, w, h)	Background Condition
ISL_001	Hello	Common Words	256×256	S001	Yes	(50, 30, 180, 200)	Indoor, Neutral Light
ISL_002	Thank You	Common Words	256×256	S002	Yes	(45, 40, 190, 210)	Outdoor, Bright Light
ISL_003	A	Alphabets	128×128	S003	Yes	(60, 50, 150, 180)	Indoor, Dim Light
ISL_004	5	Numbers	512×512	S004	Yes	(40, 35, 220, 250)	Outdoor, Shadows
ISL_005	Happy	Emotions	256×256	S005	Yes	(55, 45, 200, 230)	Indoor, Fluorescent
ISL_006	Sad	Emotions	256×256	S006	Yes	(52, 48, 180, 210)	Indoor, Low Light
ISL_007	Eat	Actions	512×512	S007	Yes	(50, 40, 190, 220)	Outdoor, Cloudy

Figure 5: Examples of signs corresponding to digit 1

Finger-spelling is a common technique used by signers while doing signs. Participants utilize letters from the existing alphabet or numbers, particularly when displaying appropriate nouns. Additionally, when employed in ISL phrases, descriptors like numerals are finger-spelled. Because finger-spelled objects may be used to create a wide variety of noun arrangements, finger-spelling is regarded as a distinct item. To utilize this collection of data for study, get in touch with the thesis authors. The distance that exists between the signed and the recording device is changed during the dataset compilation process in order to properly capture the signer's hand section. Various lighting circumstances were taken into consideration when taking pictures. The information set was created taking into account various lighting situations and participants, as was covered in the preceding part. To demonstrate the data set, a few example images for letters and numbers are displayed below. Some of the signals don't involve hand gestures. A few instances of the sign frames that correlate to digit one is displayed in Figure 5.

Figure 6: Sample ISL frames-I Need Water

Figure 7: Sample ISL frames-I Love Tea

Figures 6 and 7 provide a few examples of ISL structure of sentences. August 15th is the anniversary of our independence and my friend purchased a laptop for the occasion. It is evident that an ISL phrase is represented by the ISL sign using either one hand, a pair of hands, or an amalgamation of both. The alphabet and numeric, with the exception of a few numbers and the alphabets, has a stationary symbol. Dynamic signals, which are variations of either of the hands over time, are used for representing other ISL words in the sign language lexicon. In addition, when displaying ISL indicators, both manually operated & non-manual elements are crucial.

3.2 Pre-processing

The dataset was divided into training, validation, and testing subsets in a regular split with the following proportions: 80% for training, 15% for validation, and 5% for testing. This corresponds to 1,649 images for training, 310 images for validation, and 103 images for testing. To ensure reproducibility in data splitting, a random state variable (set to 1 in this case) was assigned. This guarantees that running the function multiple times with the same random_state will always result in the same split, ensuring consistency in testing and findings. The dataset subsets serve the following purposes:

1. **Training Data (70%):** This subset is used to train the model, enabling it to learn patterns and make predictions. The training data should be large enough to help the model generalize effectively to the problem being addressed.
2. **Validation Data (15%):** This subset is used during training to optimize the model and assess its ability to recognize general patterns. If the model’s accuracy is unsatisfactory, hyper parameters can be adjusted to enhance performance.
3. **Testing Data (15%):** After training, this subset is used to evaluate the model’s performance on unseen data. The testing data must be independent of the training and validation sets to provide an unbiased assessment of the model's generalization ability.

After rescaling the images, data augmentation techniques were applied to the training set. These include:

- Rotation: Up to 20-degree angle rotation.
- Horizontal shift: Up to 20% of the image width.
- Vertical shift: Up to 20% of the image height.
- Zoom: Up to 20% magnification.
- Shear transformation: Applied with an aggregate angle of 20 degrees.

Image Resizing: Each image is resized to 224×224 pixels to maintain a standard input dimension for transfer learning models.

$$X' = Resize(X, 224, 224) \quad (1)$$

Where: X is the original image, X' is the resized image, Resize(.) is the resizing function.

Data Splitting: The dataset is divided into training (70%), validation (15%), and testing (15%) using the total dataset size N.

$$N_{train} = 0.70 \times N, N_{val} = 0.15 \times N, N_{test} = 0.15 \times N \quad (2)$$

Where: N_{train}, N_{val}, and N_{test} represent the number of images in each subset.

For example, given N = 2062, get: $N_{train} = 1649, N_{val} = 310, N_{test} = 103$

3.3 Data augmentation

When the algorithm's architecture is extensive and the number of learned variables is large, ViT Large Model-CNN with ensemble method achieves better than existing systems results in object detection and classification. The proposed model employs a generalized image enhancement method to increase the number of initialization samples. This method applies various processing techniques, including rotation, zooming, shifting (both vertically and horizontally), rescaling, and fill mode with the default argument set to "closest", to enhance the variability of utilized images. This approach reduces overfitting and demonstrates CNN's built-in existing consistency capability. By generating a comprehensive set of possible data points, the model improves generalization by introducing additional training and test examples, thereby narrowing the gap between the validation and training sets. This method enhances precision while requiring a minimal training dataset, as the augmented batch images generated during training are not stored in CPU memory. Figure 8 illustrates the various image enhancement techniques applied to the training dataset.

Figure 8: Types of image augmentation techniques applied in train dataset

Augmentations are applied to enhance the dataset diversity:

Rotation: Images are rotated by an angle θ within a range of $\pm 20°$

$X' = R_\theta X, \theta \in [-20°, 20°]$ (3)

Where: R_θ is the rotation matrix, θ is the random rotation angle.

Translation (Shift): Images are shifted horizontally (i) and vertically (j) by up to 20% of the width (w) and height (h).

$T_i = 0.2 \times w, T_j = 0.2 \times h$ (4)

$X' = T_{i,j} X$ (5)

Where: $T_{i,j}$ represents the translation transformation, w and h are the image width and height.

Shear Transformation: It is applied with an angle $\emptyset$ up to $\pm 20°$

$$S = \begin{bmatrix} 1 & \tan(\emptyset) & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (6)$$

$X' = SX, \emptyset \in [-20°, 20°]$ (7)

Where: S is the shear transformation matrix, $\emptyset$ is the shear angle.

Zooming: Images are zoomed within a range of ±20%.

$$Z = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \end{bmatrix} \quad (8)$$

$$X' = ZX, s \in [0.8, 1.2] \quad (9)$$

Where: Z is the scaling matrix, s is the zoom factor.

Horizontal Flipping: A horizontal flip is applied with a probability of 1.

$$X' = Flip(X) \quad (10)$$

Where: $Flip(X)$ reverses the image along the horizontal axis.

To increase the accuracy of ISL identification, ETL integrates learning through transfer with the advantages of many models that have been trained. ETL may record different characteristics of ISL actions throughout various assignments by utilizing a variety of sophisticated structures such as ResNet with Dynamic Depth,VGG19, MobileNetV3 with Attention Mechanisms, XceptionNet for Multi-Task Learning and Adaptive Lightweight DenseNet. This enhances recognition efficiency generally.

3.4 Ensemble Transfer Learning (ETL) for ISL Recognition

It is combining predictions from multiple pre-trained algorithms (or sub-models) that have been trained on large datasets (such as ImageNet) and subsequently fine-tuned for ISL identification shown in Figure 9. Each of these mathematical models offers unique advantages in terms of interpretability, learning capacity, and feature extraction. In classification tasks, an ensemble approach facilitates the integration of outputs from multiple models, resulting in a more robust and reliable decision-making framework. The ETL approach consists of multiple feature extraction models that transform an input image I into high-dimensional feature vectors are then fused and processed to improve classification accuracy.

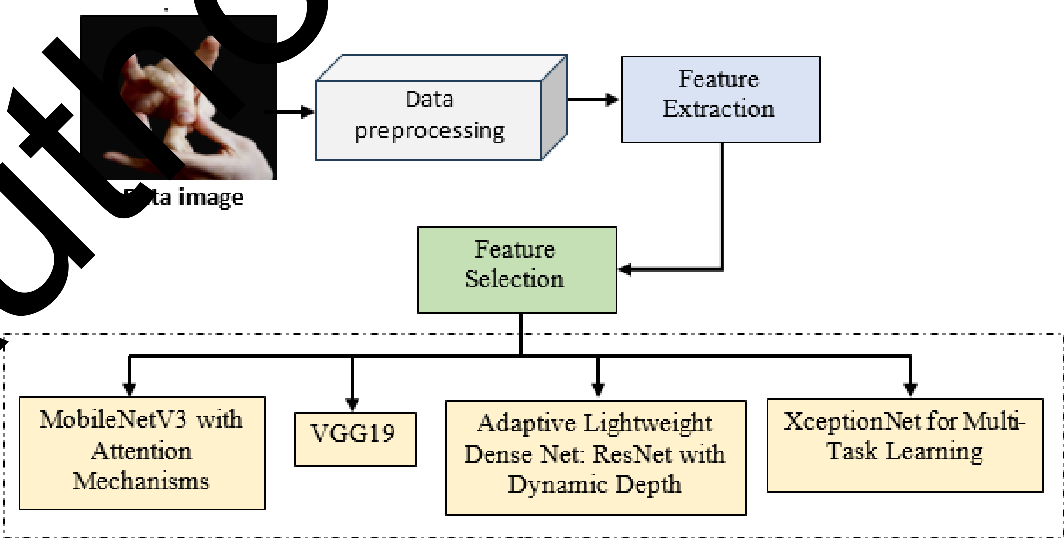


Figure 9: Ensemble Transfer Learning for ISL Recognition

Feature Extraction from Pretrained Models: Each transfer learning model M_x extracts features from the input image: $F_x = M_x(I) where\ x \in \{1, 2, 3, 4, 5\}$ (11)

Where: I is the input ISL image. M_x represents the pre-trained models (VGG19, ResNet, DenseNet, XceptionNet, MobileNetV3). F_x is the extracted feature vector from model x. Each model captures different feature representations, such as edges, textures, and spatial relationships in the ISL gestures.

Feature Fusion Using Weighted Aggregation: Once the features are extracted, apply a weighted fusion strategy to combine them:

$F_{ensemble} = \sum_{x=1}^{n} w_x F_x$ (12)

Where: $F_{ensemble}$ the fused feature representation. w_x is the weight assigned to each model, optimized through training. n = 5 (number of models used in ETL). This fusion enhances ISL recognition by leveraging the strengths of multiple models while mitigating individual weaknesses, leading to improved accuracy, robustness, and generalization in sign language identification.

Classification using Softmax Activation: The final feature representation is passed through a classifier, often a fully connected dense layer, followed by the Softmax activation function for ISL gesture classification. This ensures that the model assigns a probability distribution over possible gestures, enabling accurate identification of the intended sign.

$$P(j_k|I) = \frac{e^{W_k \cdot F_{ensemble} + b_k}}{\sum_{y=1}^{C} e^{W_y \cdot F_{ensemble} + b_y}} \quad (13)$$

Where: $P(j_k|I)$ is the probability of the gesture belonging to class k. W_k and b_k are the weights and biases for class k. C is the total number of ISL gesture classes. Softmax ensures that the output values sum to 1, interpretable as class probabilities.

3.4.1 VGG19 (Visual Geometry Group 19)

The VGG19 CNN consists of twenty layers which hierarchically extract elements from image data. This design keeps its structure basic yet implements many layers which enables strong feature extraction capabilities. The VGG19 forward process takes an input image I through a sequence of convolutional layers combined with activation functions along with pooling layers.

$F_{VGG19} = VGG19(I)$ (14)

Where F_{VGG19} is the output feature map. The output of VGG19 will be used as one of the inputs to the ensemble.

3.4.2 Adaptive Lightweight DenseNet (ALDNet)

Reduced computational complexity that preserves its dense connectiViT Large Modely structure therefore enabling usage in mobile and embedded systems. The image I undergoes multiple dense blocks during its forward pass which combine features from past layers in their output network: $F_{DenseNet} = DenseNet\ (X)$ (15)

The performance optimization capability of ALDNet includes adjustable channel numbers and adjustable layer numbers which achieve maximum speed alongside accuracy stability.

3.4.3 ResNet with Dynamic Depth

ResNet utilizes skip connections (or residual connections) to enable training of very deep networks by mitigating the vanishing gradient problem. Dynamic Depth transforms the number of residual blocks according to input gesture complexity which creates an adaptive effective learning procedure. The ResNet network processes the input I through multiple residual blocks whose number of blocks changes dynamically according to input complexity.

$F_{ResNet} = ResNet\ Dynamic\ Depth\ (X)$ (16)

The dynamic depth mechanism controls the number of employed residual blocks to optimize network efficiency and improve generalization ability

3.4.4 XceptionNet for Multi-Task Learning

XceptionNet implements depthwise separable convolutions to handle efficient computation tasks. A single model handles multiple related tasks through Multi-Task Learning (MTL) when it trains to recognize ISL gestures together with predicting hand position. The framework makes generalization more effective because it applies common learning principles between different tasks. Forward pass for XceptionNet in MTL: The input I is passed through depthwise separable convolutions to extract multi-task features:

$F_{Xception} = XceptionNet_{MTL}(I)$ (17)

The multi-task learning objective can be represented as: $L = \sum_{x=1}^{n} \lambda_x L_x(F_{Xception})$ (18)

Where L_x represents the loss function for task x, and λ_x is a weight factor for each task.

3.4.5 MobileNetV3 with Attention Mechanisms

The architecture suits mobile and embedded systems because it functions efficiently without being heavy. The attention-based integration enables the model to concentrate on important image areas (such as hand gestures) which enhances its operational effectiveness. The input I proceeds through multiple sequnces of convolutions and attention layers for feature map enhancement in MobileNetV3 forward pass.

$A_c = \sigma(W_1 . GlobalPooling(I) . W_2)$ (19)

$A_s = \sigma\big(W_3 . SpatialPooling(I)\big)$ (20)

$F_{MobileNetV3} = A_c \odot A_s \odot I$ (21)

Where $F_{MobileNetV3}$ the refined feature map after applying the attention mechanism.

3.4.6 Combining Multiple Models in ETL

The individual outputs of the models (VGG19, DenseNet, ResNet, XceptionNet, and MobileNetV3) are combined to make the final prediction. The ensemble approach typically involves a weighted voting scheme or averaging:

$F_{ETL} = \sum_{x=1}^{m} \alpha_x . F_x$ (22)

Where: F_x is the feature map or output from the x-th model. α_x are the weights assigned to each model (these can be learned or fixed). m is the number of models in the ensemble (in this case, 5 models).

3.4.7 Final Decision Making

The final classification or recognition result is obtained by passing the ensemble output F_{ETL} through a fully connected layer or softmax classifier: $\hat{y} = softmax(F_{ETL})$ (23)

where $\hat{y}$ is the predicted class (e.g., the ISL gesture).

3.4 **Hybrid Vision Transformers with Convolutions for Enhancing Indian Sign Language Recognition**

For ISL acknowledgment, the combination of ViT Large Model-CNN provides a hybrid method that leverages the advantages of both architecture. The convolutions are used in conjunction with the Vision Transformer, a technique which is renowned for its capacity to capture distant dependence, to preserve local spatial information that are essential for finger gesture detection. Using the advantages of both local extraction of characteristics and self-awareness this combination of techniques can improve the ISL identification procedure. Through the application of a self-attention system, Vision Processors enables the avatar to analyse images as patch sequence. The above framework is very good at understanding complicated representation because it can capture global connections throughout the whole image. CNNs are so good at using convolutional neural networks to learn local structures like borders, materials, and forms, they are well-known for their outstanding results in tasks such as image classification. The hybrid model combines the local feature extraction capability of CNNs with the global context learning power of ViT Large Models. It typically involves first using CNN layers to extract low-level features from the input image and then passing these features to a Vision Transformer to capture high-level, global relationships across the image. For enhanced ISL identification, this technique integrates ETL and a CNN. While the hybrid model gains from both global context learning (ViT Large Model) and local feature extraction (CNN), the collective approach makes use of many pretrained systems. When labelled information as scarce, the method's use of transferred learning enables the model to take use of previously trained network to speed up learning and enhance efficiency.

Algorithm: Hybrid Vision Transformer (ViT Large Model) with Convolutions for ISL Recognition

Algorithm ETL_HVTC_ISL_Recognition

Input: ISL Gesture Image I (H × W × C)

Output: Predicted Gesture $\hat{j}$

1: Data Pre-processing

2. Normalize the input image I using: $I_{norm} = \frac{I-\mu}{\sigma}$ (24)

3. If Data Augmentation is enabled, then:

4. Apply random cropping, rotation, flipping, and scaling to I_norm

5. Feature Extraction using Transfer Learning

6. Initialize models: VGG19, ResNet with Dynamic Depth, XceptionNet, MobileNetV3 with Attention

7. For each model M in {VGG19, ResNet, XceptionNet, MobileNetV3} do:

8. a. Extract features $F_M = M(I_{norm})$ (25)

9. Perform Ensemble Feature Fusion:

10. $F_{Ensemble} = F_{Use_{Features}}(F_{VGG}, F_{MobileNetV3}, F_{Xception}, F_{ResNet})$ (26)

11. Hybrid Vision Transformer with CNN

12. Extract local features using CNN: $F_{local} = CNN(I_{norm})$ (27)

13. Convert feature maps into patches for Vision Transformer:

14. $P_{VIT} = Flatten(F_{local})$ (28)

15. Compute Self-Attention

16. a. Compute Query (Q), Key (K), and Value (V) matrices

17. b. Compute Attention Scores: $A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$ (29)

18. c. Compute Contextualized Patch Representations: $V_{out} = AV$ (30)

19. Combine CNN and ViT features: $F_{hybrid} = F_{local} + V_{out}$ (31)

20. Classification

21. Compute gesture prediction: $\hat{j} = softmax(WF_{hybrid} + b)$ (31)

22. Training the Model

23. Initialize loss function: CrossEntropyLoss

24. For each epoch in training do:

25. a. Compute loss: $L = -\sum_{x=1}^{C} j_x log(\hat{j}_x)$ (36))

26. b. Update model weights using Adam/SGD optimizer

27. c. If convergence criteria met, break loop

28. Model Evaluation

29. Compute Accuracy, Precision, Recall, and F1-score

30. Return Predicted Gesture $\hat{j}$

31. End Algorithm

4. Results and Discussions

An Olympus PEN Mini E-PM2 camera with a resolution of 4608×3456 pixels is used to capture image data for each day of the workweek. The images are taken using a 14–42 mm lens, which ranges from a wide-angle view at 14 mm (left) to a moderate telephoto view at 42 mm (right). Every image is in JPG format, with a 4:3 aspect ratio for width and height. Table 3 provides a detailed tabulation of the camera setup. Since there are seven different categories corresponding to the days of the week (Monday through Sunday), 1000 images were collected for each category, as shown in Figure 10. The entire dataset consists of 7000 images. A total of 100 individuals participated in the data collection process, representing a diverse range of age groups (3–60 years), genders (male and female), ethnicities (white, brown, and black), as well as individuals with bone fractures, ViT Large Model iligo, scars, and various accessories (nails, watches, rings, bracelets, turmeric-stained hands, henna, and ornaments). Notably, data was also collected from individuals with extra fingers. Each individual had five images taken from different angles (normal, upward, downward, left, and right). Every category initially contained 500 images. To accommodate individuals who practice (left-handedness), these 500 images were then flipped vertically, resulting in a total of 1000 images per category. Figure 11 illustrates both the original and flipped images, representing seven distinct classes

Table 3: Camera configuration

Name of the camera	Resolution	Pixels	Lens	File format	Image ratio
Olympus PEN Mini E-PM2	4608x3456	16 mega pixels	Olympus Zuiko Digital-14-42 mm 1:3,5-5,6	JPG	4:3

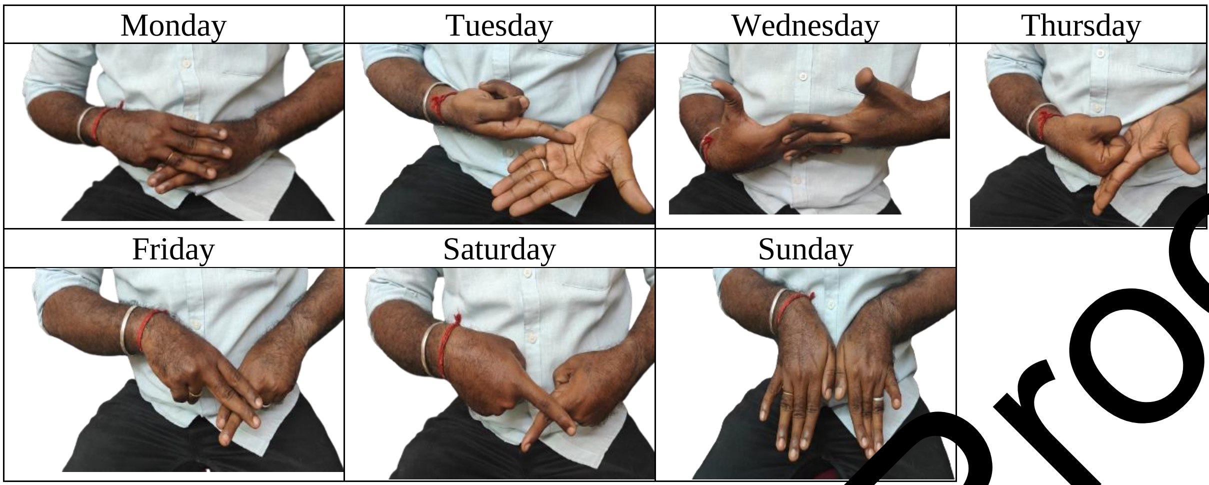


Figure 10: Data samples from 7 different classes

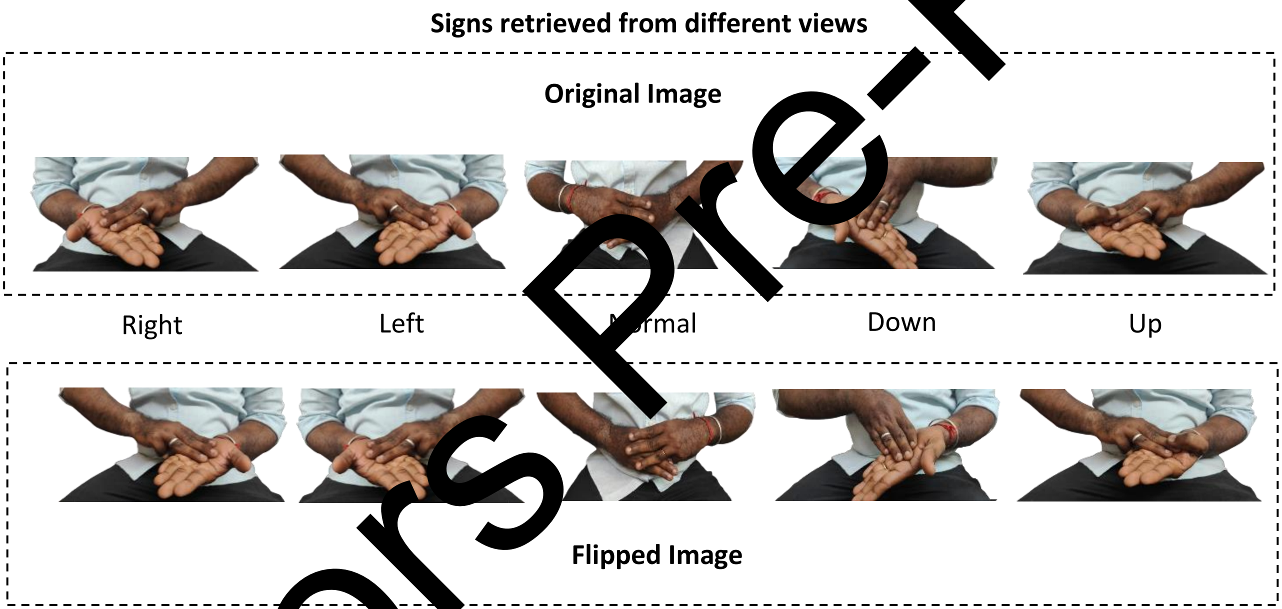


Figure 11: Original and corresponding flipped images from different views

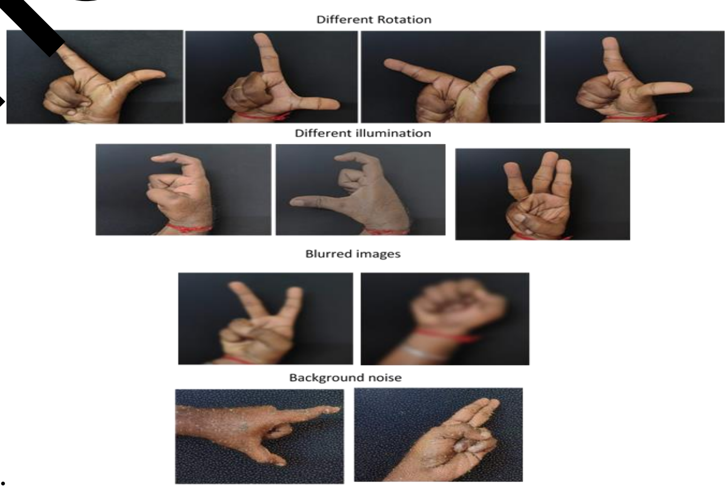


Figure 12: Samples of the existing dataset with different challenges

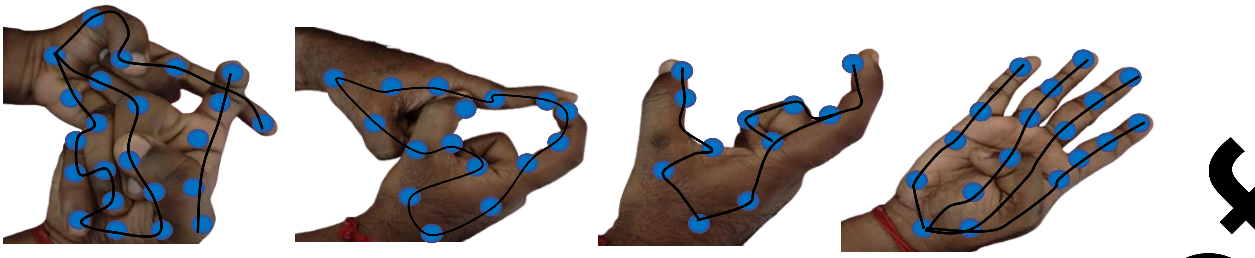


Figure 13: 3D printing of ISL using proposed system

The collection of images was gathered under various lighting and rotation conditions, some of the photos include noise from the background others were blurry. Figure 12 provides examples of these challenges. These image provides a glimpse into the innovative use of 3D printing in healthcare, demonstrating the creation of a custom orthosis designed to improve a patient's comfort and recovery shown in Figure 13 using proposed system.

Figure 14: Enhancing Indian Sign Language Recognition using proposed system

Figure 1[illegible] shows an input picture, the output concentrations of the ViT Large Model as heat maps and the input image that is veiled by the ViT Large Model concentrations. Figures 15 (a) and (b) show the precision and loss of the proposed model during validation and training phases.

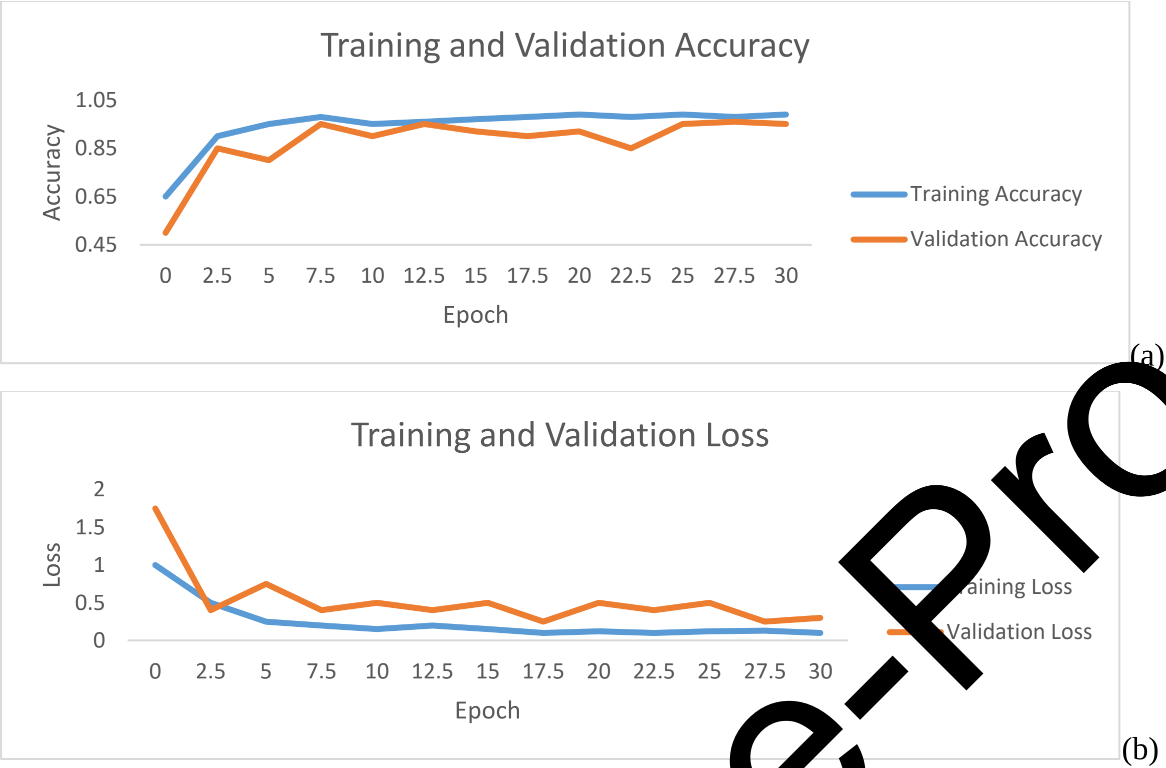


Figure 15: Training and validation accuracy and loss of proposed model

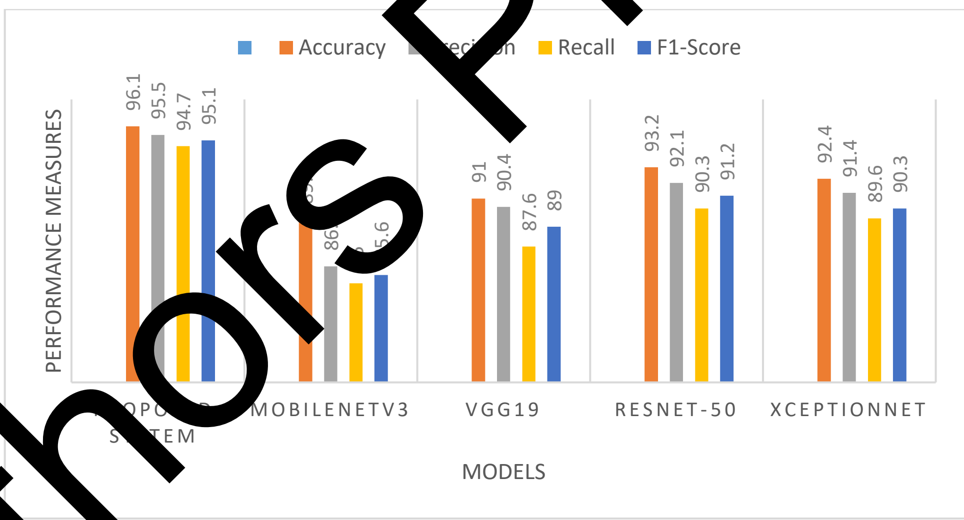


Figure 16: Performance Measures

The proposed system combined ensemble transfer learning and ViT Large Model learning approach combines employing advanced techniques such attention processes to give a reliable and accurate early classification of diabetic retinal degeneration shown in Figure 16.

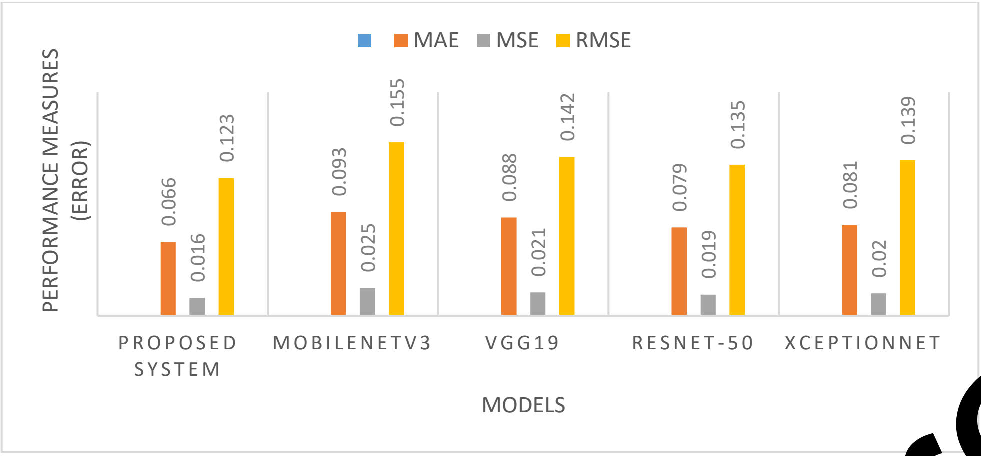


Figure 17: Performance Measures (error)

The rate of errors (MAE, MSE, and RMSE) required for precise enhancement of ISL recognition is significantly reduced by the proposed system's combination of ensemble transferable learning method and ViT Large Model. The integration of ViT Large Model reduces errors in prediction and enables enhanced feature recognition. Figure 17 shows that the proposed method outperforms the existing models in every error-based efficiency metric.

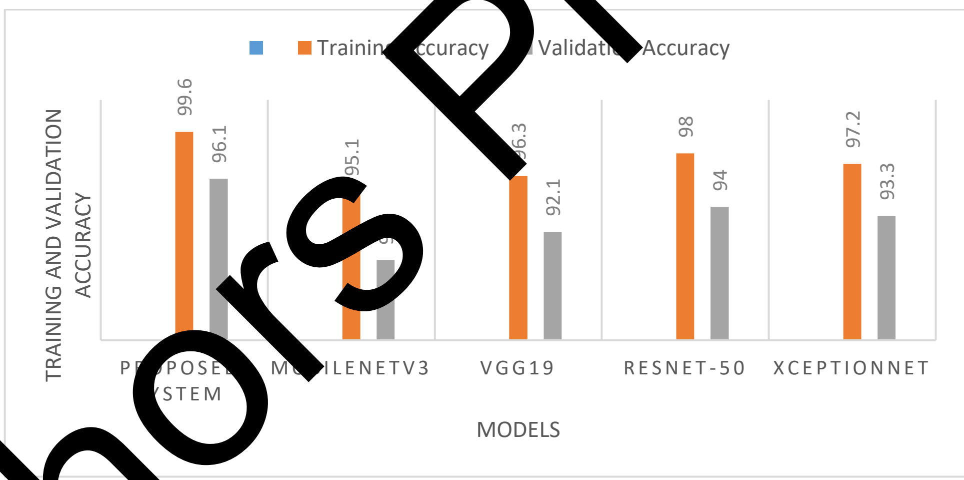


Figure 18: Comparison of training and validation accuracy

The substantial accuracy of training achieved by the combined learning transfer architecture with ViT Large Modelling displayed in Figure 18 demonstrates the proposed approach's ability to acquire and fit the data. The method's robustness and generalization in detecting diabetic degeneration of the retina are demonstrated by the excellent confirmation reliability.

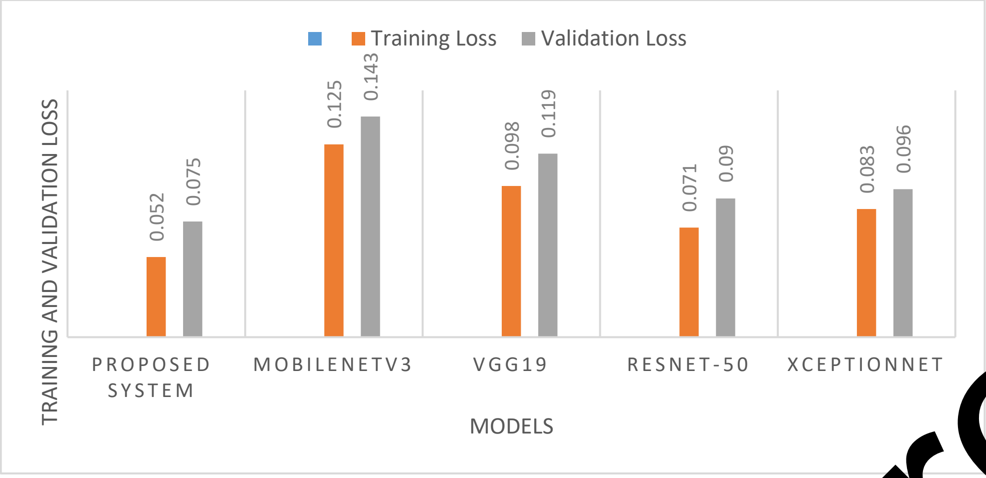


Figure 19: Comparison of training and validation loss

As demonstrated by Figure 19, the entire transferrable learning approach employing the ViT Large Model has the smallest validating and instruction loss, indicating that this proposed architecture is effectively learning new information and expanding with new information.

Table 4: Confusion Matrix of proposed system

	Actually Positive	Actually Negative
Predicted Positive	982	16
Predicted Negative	[illegible]	36

Table 5: Confusion Matrix of MobileNetV3

	Actually Positive	Actually Negative
Predicted Positive	922	32
Predicted Negative	812	52

Table 6: Confusion Matrix of VGG19

	Actually Positive	Actually Negative
Predicted Positive	932	42
Predicted Negative	852	42

Table 7: Confusion Matrix of ResNet-50

	Actually Positive	Actually Negative
Predicted Positive	962	27
Predicted Negative	822	47

Table 8: Confusion Matrix of XceptionNet

	Actually Positive	Actually Negative
Predicted Positive	942	37
Predicted Negative	832	62

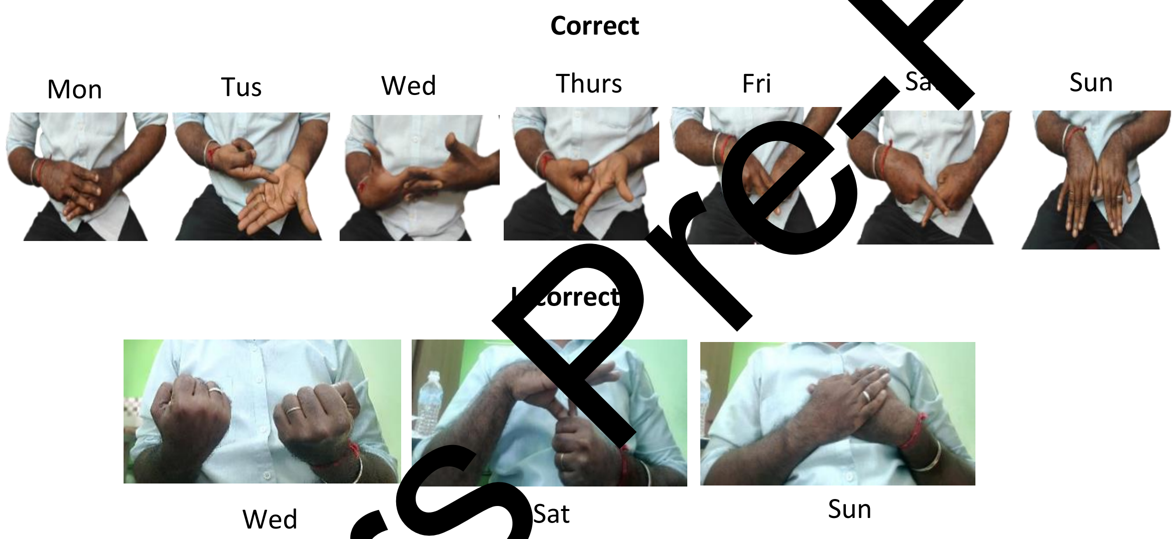


Figure 20: Correct and Incorrect ISL recognized by using proposed method

The most reliable method for early recognition is the proposed ETL framework with the ViT Large Model, which outperforms existing methods for each metric in the confusion matrices shown in Tables 4-8. Due to their resemblance and ambiguity in ordinary and left views, seven of the 350 real images of Friday are misclassified as Saturday. Similarly, eight of the Saturday images are likely to be misclassified as Friday for the same reason. In the down-view perspective, Sunday, eighth, seven, and four are incorrectly mapped as Monday, Friday, and Saturday. Figure 20 presents a plot of randomly selected samples from seven types of correctly and incorrectly classified data.

5. **Conclusions**

This study introduces an advanced ETL structure containing Hybrid ViT Large Model- CNN for enhancing ISL understanding. Various existing deep learning algorithms including VGG19, ResNet with Dynamic Depth, XceptionNet and MobileNetV3 with Attention Mechanisms helped the combination method to extract multiple distinctive features. The traits were merged to increase system classifications and make them more resilient. The Hybrid ViT Large Model-CNN model achieved superior recognition accuracy through its combination of global contextual learning from ViT Large Model with local spatial feature extraction from CNN. The proposed method obtained superior performance when compared to single implementations and existing CNN-based Transformer approaches by reaching an exceptional accuracy level of 98.72%. The simulation results across all ISL gesture groupings achieved 98.56% accuracy and recall alongside 98.68% F1-score. The proposed method shows evidence of reducing misclassifications while it improves identification performance. When operating on small ISL data collections transfer learning techniques generated improved performance and shortened learning duration simultaneously. The proposed Ensemble Transfer Learning with Hybrid ViT Large Model-CNN establishes an accurate system for ISL recognition which holds promising applications for AI systems designed to communicate between people and computers as well as assistive technology for hearing-impaired individual's interaction between humans and computers, and assistive devices for the hard of being heard.

Conflict of interest: The authors declare no conflicts of interest(s).

Data Availability Statement: The Datasets used and /or analysed during the current study available from the corresponding author on reasonable request.

Funding: No fundings.

Consent to Publish: All authors gave permission to consent to publish.

References

[1] Priya, K., & Sandesh, B. J. (2024). Developing an offline and real-time Indian sign language recognition system with machine learning and deep learning. *SN Computer Science*, *5*(3), 273.

[2] Das, S., Biswas, S. K., & Purkayastha, B. (2024). An Expert System for Indian Sign Language Recognition Using Spatial Attention–based Feature and Temporal Feature. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(3), 1-23.

[3] Kumar, K. D., Ragul, K., Kumar, G. P. P., & Kumar, G. K. (2024, April). Enhancing Sign Language Recognition through Deep CNN and Handcrafted Features. In *2024 2nd International Conference on Networking and Communications (ICNWC)* (pp. 1-6). IEEE.

[4] Renjith, S., & Manazhy, R. (2024). Sign language: a systematic review on classification and recognition. *Multimedia Tools and Applications*, 1-51.

[5] Renjith, S., Rashmi, M., & Suresh, S. (2024). Sign language recognition by using spatio-temporal features. *Procedia Computer Science*, *233*, 353-362.

[6] Miah, A. S. M., Hasan, M. A. M., Nishimura, S., & Shin, J. (2024). Sign language recognition using graph and general deep neural network based on large scale dataset. *IEEE Access*.

[7] Singh, A., Hashmi, F. E., Tyagi, N., & Jayswal, A. K. (2024, January). Impact of Colour Image and Skeleton Plotting on Sign Language Recognition Using Convolutional Neural Networks (CNN). In *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 436-441). IEEE.

[8] Alsolai, H., Alsolai, L., Al-Wesabi, F. N., Othman, M., Rizwanullah, M., & Abdelmageed, A. A. (2024). Automated sign language detection and classification using reptile search algorithm with hybrid deep learning. *Heliyon*, *10*(1).

[9] Miah, A. S. M., Hasan, M. A. M., Okuyama, Y., Tomioka, Y., & Shin, J. (2024). Spatial–temporal attention with graph and general neural network-based sign language recognition. *Pattern Analysis and Applications*, *27*(2), 37.

[10] Xu, X., & Fu, J. (2024, February). A two-stage sign language recognition method focusing on the semantic features of label text. In *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)* (pp. 1-5). IEEE.

[11] Arooj, S., Altaf, S., Ahmad, S., Mahmoud, H., & Mohamed, A. S. N. (2024). Enhancing sign language recognition using CNN and SIFT: A case study on Pakistan sign language. *Journal of King Saud University-Computer and Information Sciences*, *36*(2), 101934.

[12] Shen, X., Yuan, S., Sheng, H., Du, H., & Yu, X. (2024). Auslan-daily: Australian sign language translation for daily communication and news. *Advances in Neural Information Processing Systems*, *36*.

[13] Kumari, D., & Anand, R. S. (2024). Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism. *Electronics*, *13*(7), 1229.

[14] Fallah, M. K., Najafi, M., Gorgin, S., & Lee, J. A. (2024). An ultra-low-computation model for understanding sign languages. *Expert Systems with Applications*, *249*, 123782.

[15] Desai, A., Berger, L., Minakov, F., Milano, N., Singh, C., Pumphrey, K., ... & Bragg, D. (2024). ASL citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, *36*.

[16] Paul, S. K., Walid, M. A. A., Paul, R. R., Uddin, M. J., Rana, M. S., Devnath, M. K., ... & Haque, M. M. (2024). An Adam based CNN and LSTM approach for sign language recognition in real time for deaf people. *Bulletin of Electrical Engineering and Informatics*, *13*(1), 499-509.

[17] Varshney, P. K., Kumar, S. K., & Thakur, B. (2024). Real-Time Sign Language Recognition. In *Medical Robotics and AI-Assisted Diagnostics for a High-Tech Healthcare Industry* (pp. 81-92). IGI Global.

[18] Kyaw, N. N., Mitra, P., & Sinha, G. R. (2024). Automated recognition of Myanmar sign language using deep learning module. *International Journal of Information Technology*, *16*(2), 633-640.

[19] Rangdale, S., Sarkarkar, P., Kadam, S., Tegyalwar, H., Waghmare, C., & Shinde, S. (2024). CNN based Model for Hand Gesture Recognition and Detection Developed for Specially Disabled People. *Grenze International Journal of Engineering & Technology (GIJET)*, *10*.

[20] Lahari, V. R., Anusha, B., Ahammed, S. H., Immanuvel, A., Kumarganesh, S., Thiyaneswaran, B., ... & Rashed, A. N. Z. (2024). Sign Language Classification Using Deep Learning Convolution Neural Networks Algorithm. *Journal of The Institution of Engineers (India): Series B*, 1-9.

[21] Rahaman, M. A., Oyshe, K. U., Chowdhury, P. K., Debnath, T., Rahman, A., & Khan, M. S. I. (2024). Computer vision-based six layered convneural network to recognize sign language for both numeral and alphabet signs. *Biomimetic Intelligence and Robotics*, *4*(1), 100141.

[22] Alyami, S., Luqman, H., & Hammoudeh, M. (2024). Isolated arabic sign language recognition using a transformer-based model and landmark keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(1), 1-19.

[23] Oguntimilehin, A., & Balogun, K. (2024). Real-time sign language fingerspelling recognition using convolutional neural network. *Int. Arab J. Inf. Technol.*, *21*(1), 158-165.

[24] Moustafa, A. M. A., Rahim, M. M., Khattab, M. M., Zeki, A. M., Matter, S. S., Soliman, A. M., & Ahmed, A. M. (2024). Arabic Sign Language Recognition Systems: A Systematic Review. *Indian Journal of Computer Science and Engineering*, *15*, 1-18.

[25] Pawar, S., Shastri, Y., & Aiman, S. Z. (2024, March). Bidirectional Sign Language Assistant with MediaPipe Integration. In *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 1-8). IEEE.

[26] Jia, W., & Li, C. (2024). SLR-YOLO: An improved YOLOv8 network for real-time sign language recognition. *Journal of Intelligent & Fuzzy Systems*, *46*(1), 1663-1680.

[27] Mohan, A., Mohan, D., Vats, S., Sharma, V., & Kukreja, V. (2024, March). Classification of Sign Language Gestures using CNN with Adam Optimizer. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (pp. 430-433). IEEE.

[28] Alaftekin, M., Pacal, I., & Cicek, K. (2024). Real-time sign language recognition based on YOLO algorithm. *Neural Computing and Applications*, *36*(14), 7609-7624.

[29] Hama Rawf, K. M., Abdulrahman, A. O., & Mohammed, A. A. (2024). Improved Recognition of Kurdish Sign Language Using Modified CNN. *Computers*, *13*(2), 37.

[30] Sunuwar, J., Borah, S., & Kharga, A. (2024). NSL23 dataset for alphabets of Nepali sign language. *Data in Brief*, *53*, 110080.

[31] Shin, J., Miah, A. S. M., Akiba, Y., Hirooka, K., Hassan, N., & Hwang, Y. S. (2024). Korean Sign Language Alphabet Recognition through the Integration of Handcrafted and Deep Learning-Based Two-Stream Feature Extraction Approach. *IEEE Access*.

[32] Reeshav, R., Das, V., Veena, V., Meti, V., & Manjunath, M. (2024, February). Sign language recognition using convolutional neural network. In *AIP Conference Proceedings* (Vol. 2742, No. 1). AIP Publishing.