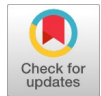


AI-Driven Indian Sign Language Recognition Using Hybrid CNN-BiLSTM Architecture for Divyangjan

Nithyanandh S



Abstract: Barriers in expressive and receptive communication for individuals with hearing and speech impairments remain a significant challenge to achieving equitable social interaction and digital accessibility. These limitations restrict their participation in everyday conversations, education, and professional environments, emphasizing the urgent need for intelligent assistive communication technologies. This research introduces an AI-driven real-time Indian Sign Language (ISL) recognition framework that integrates advanced computer vision and deep learning techniques to translate hand gestures into textual outputs. The primary objective of this study is to design a lightweight, accurate, and real-time sign language translator suitable for Divyangjan, ensuring inclusivity in educational and social interactions. The proposed research employs Mediapipe-based hand landmark extraction to detect 21 key points in each gesture frame, followed by preprocessing and normalisation to create robust spatial representations. A hybrid Convolutional Neural Network processes these landmark vectors through a Bidirectional Long Short-Term Memory (CNN-BiLSTM) model that captures both spatial and temporal dependencies in gesture motion, allowing it to recognise static and dynamic gestures such as "J" and "Z." The system was trained on a self-collected dataset of over 26,000 gesture images covering all 26 ISL alphabets. Experimental analysis demonstrates that the proposed model achieved 97.8% accuracy, 97.4% precision, 97.2% recall, 97.3% F1-score, and a validation loss of only 0.08, outperforming traditional classifiers such as Random Forest and SVM by a significant margin. The trained model performs robustly under varying lighting, background, and hand orientation conditions, ensuring high reliability for real-world deployment. The novelty of this study lies in the fusion of Mediapipe landmark extraction with a temporal deep learning framework for continuous ISL gesture translation within an interactive CustomTkinter GUI. This human-centric, computationally efficient design enables accessible, real-time communication for the deaf and hard-of-hearing community, contributing to socially assistive AI systems that promote digital inclusivity and empowerment for Divyangjan.

Keywords: Artificial Intelligence, Bidirectional Long Short-Term Memory (BiLSTM), Convolutional Neural Network (CNN), Gesture Recognition, Indian Sign Language (ISL).

Nomenclature:

CNNs: Convolutional Neural Networks

BiLSTM: Bidirectional Long Short-Term Memory

AI: Artificial Intelligence

ISL: Indian Sign Language

SLR: Sign Language Recognition

ARSL: Arabic Sign Language

ASL: American Sign Language

IRSG: Institutional Research Seed Grant

I. INTRODUCTION

In an increasingly digital and interconnected world, communication remains a fundamental bridge that enables inclusion, education, and social participation. However, this bridge often remains incomplete for individuals with hearing and speech impairments, collectively recognized as *Divyangjan*, creating barriers to equal access in education, employment, and daily interactions. Sign language is their primary mode of expression, yet its understanding among the general population is limited. This disparity underscores the pressing need for intelligent, accessible technologies to translate sign language into textual or spoken forms, facilitating seamless communication between hearing and non-hearing individuals [1].

Recent advancements in artificial intelligence (AI) and computer vision have transformed the landscape of assistive communication tools. AI-based sign language recognition (SLR) systems use machine learning models to identify hand gestures, shapes, and movements captured by cameras and translate them into corresponding text or speech in real time [2]. These systems empower the *Divyangjan* community and contribute to inclusive digital ecosystems envisioned under global sustainable development frameworks, particularly SDG-10 (Reduced Inequalities) and SDG-4 (Quality Education). Traditional methods relying on glove-based sensors and static image matching have shown limitations in handling natural, dynamic gestures, prompting researchers to adopt deep learning frameworks capable of understanding complex spatial and temporal dependencies in hand movements [3].

Integrating Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) models has emerged as a promising direction for accurate gesture classification. CNNs extract spatial hierarchies such as finger positions and palm orientation, while BiLSTM layers capture motion continuity across gesture sequences [4]. This dual capability enables robust recognition of dynamic gestures like "J" and "Z" that involve temporal motion, which earlier systems often misclassified. Furthermore, real-time gesture-detection frameworks based on Mediapipe and OpenCV provide reliable landmark extraction,

Manuscript received on 26 October 2025 | First Revised Manuscript received on 04 November 2025 | Second Revised Manuscript received on 20 November 2025 | Manuscript Accepted on 15 December 2025 | Manuscript published on 30 December 2025.

*Correspondence Author(s)

Dr. Nithyanandh S*, Department of MCA, PSG College of Arts & Science, Coimbatore, (Tamil Nadu), India. Email: nknithu@gmail.com. ORCID ID: [0000-0003-4779-7339](https://orcid.org/0000-0003-4779-7339)

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

significantly improving performance across varying lighting and background conditions [5].

Although many studies have explored sign language recognition, most focus on American or Arabic sign languages. At the same time, Indian Sign Language (ISL) remains comparatively underrepresented despite India's large hearing-impaired population. Given ISL's linguistic and regional variations, developing an adaptable, lightweight recognition framework is critical. Addressing this gap, the proposed study introduces an AI-driven ISL recognition model that employs Mediapipe-based hand landmark extraction, combined with a hybrid CNN-BiLSTM network, to translate gestures into text via a responsive CustomTkinter graphical interface [6]. This end-to-end framework captures real-time gestures from a standard webcam, processes 21 key hand landmarks per frame, normalizes the features, and predicts alphabetic or functional gestures with high precision.

The relevance of this research lies not only in its technological novelty but also in its social impact. A low-cost, real-time, and highly accurate translation mechanism eliminates the need for specialised equipment or interpreters, making digital communication more accessible to Divyangjan users [7].

The approach aligns with India's *Accessible Bharat* initiative, promoting equitable participation through assistive AI systems. An experimental evaluation on a self-curated dataset comprising 26,000 gesture images demonstrates exceptional performance and surpasses that of conventional machine learning classifiers [8].

Moreover, this study contributes to human-computer interaction research by integrating context-aware AI into accessibility applications. The hybrid model enhances recognition reliability by learning subtle inter-finger dependencies and motion variations, which are often overlooked in static frame-based methods [9]. Unlike earlier works that focused solely on classification accuracy, the proposed system emphasises real-world usability, ensuring smooth performance on consumer-grade hardware. This design principle is particularly valuable for educational institutions, rehabilitation centres, and public service platforms that support deaf individuals.

The proposed deep learning framework is a technological enabler bridging the communication divide between hearing and non-hearing communities. Its deployment in mobile or desktop environments can democratize access to inclusive communication, aligning innovation with compassion. Thus, the proposed model is a significant step toward realising AI for social good, empowering the Divyangjan sector through intelligent, affordable, and human-centric solutions [10].

II. LITERATURE REVIEW

The growing advancements in artificial intelligence and deep learning have significantly influenced the development of assistive technologies for the hearing-impaired community. Sign language recognition (SLR) systems are increasingly designed using data-driven frameworks that capture the spatial and temporal attributes of hand gestures, finger orientations, and motion trajectories. The literature survey reflects a significant shift from handcrafted feature extraction methods toward deep learning models that learn representations directly from visual data. This section

reviews recent studies on SLR technologies, analyzing their methodologies, architectures, limitations, and how these drawbacks form the foundation for the proposed hybrid CNN-BiLSTM framework.

A. Deep Learning Architectures for Sign Language Recognition

Researchers have recently explored multiple architectures to enhance gesture classification accuracy and model generalization. A notable study introduced an attention-based inception and BiLSTM architecture to classify Arabic Sign Language (ARSL) using real-time video sequences [11]. The system demonstrated superior temporal understanding and spatial context by combining convolutional layers and recurrent units. However, it had high computational requirements during training, limiting its suitability for low-end hardware commonly used in assistive applications.

A follow-up study proposed a hybrid CNN-LSTM model for continuous ARSL recognition, achieving greater stability than on isolated gesture datasets [12]. While it addressed motion continuity, it relied heavily on large, annotated datasets, which are scarce for regional sign languages like ISL. The lack of diversity in training data reduced adaptability to multiple skin tones and hand orientations. Similarly, another experiment using multi-view hand-skeleton models captured with RGB cameras improved depth perception but failed to maintain consistency under changing lighting conditions [13].

An investigation is conducted using a skeleton-aware multimodal network that combines pose estimation and optical flow features to interpret dynamic hand sequences [14]. Despite its ability to model body kinematics, it struggled with finger-level recognition precision. The high-dimensional data required substantial GPU memory, making real-time processing challenging. In contrast, word-level recognition models explored by several researchers have introduced deep CNN frameworks that use video inputs to identify entire sign words rather than individual letters [15]. These models required extensive temporal data, often unavailable for isolated sign alphabets.

B. Review of Hybrid and Attention-Based Models

Integrating attention mechanisms in hybrid networks improved focus on relevant spatiotemporal regions, enhancing classification performance. Reviews from 2023 identified the rise of transformer-based and attention-guided CNN models for continuous sign interpretation [16]. However, these transformer variants required large-scale training corpora and powerful GPUs, limiting deployment on resource-constrained platforms. Although attention layers increased interpretability, they also introduced overfitting risks in smaller datasets such as ISL, where gesture diversity is limited.

A comprehensive analysis of the evolution of deep SLR frameworks emphasized the potential of hybrid CNN-RNN models and outlined their challenges in learning non-linear hand trajectories and maintaining temporal coherence across frames [17]. The review suggested integrating graph-based learning or sequential modelling to improve temporal modelling. Despite

these advancements, real-time systems still faced latency issues due to complex feature-extraction pipelines.

A detailed study of a hybrid DNN for recognising Indian Sign Language gestures related to medical emergencies were conducted [18]. The system employed a convolutional feature extractor and a multilayer perceptron classifier, achieving acceptable accuracy under controlled laboratory conditions. However, the model lacked resilience to environmental variations, particularly in uncontrolled lighting or background clutter. This limitation highlighted the need for lightweight models that balance performance and computational efficiency.

Another line of research investigated self-supervised and transformer-based pretraining techniques for SLR. One study introduced a hand-model-aware self-supervised learning system to enhance spatial feature encoding through unlabelled data [19]. Though promising, such pretraining approaches required extensive computational resources, limiting their adoption in low-power embedded systems or assistive devices.

A complementary effort proposed a video transformer model that processed temporal relationships without recurrent structures [20]. It achieved competitive results in benchmark datasets but showed slower inference in real-time streaming, an essential requirement for interactive communication tools.

C. Sensor-Based and Vision-Based Systems

Beyond deep learning, traditional sensor-based recognition techniques employed gloves or wearable sensors to capture finger flexion and orientation data. Although early glove-based systems demonstrated high precision, their dependence on physical hardware restricted usability and convenience for Divyangjan individuals. In contrast, vision-based approaches using standard cameras and frameworks such as MediaPipe provided contactless solutions with lower maintenance requirements. A comprehensive 2021 review analysed both types of systems and concluded that deep learning-based vision systems offer greater scalability and accessibility [21].

Recent innovations in keypoint tracking and pose estimation have further transformed gesture interpretation. For instance, a 2025 study used Media-Pipe keypoint tracking integrated with deep CNN layers for American Sign Language (ASL) recognition, achieving near-human accuracy [22].

However, even with improved precision, the approach demanded high-quality camera input and often struggled with occlusions or partial hand visibility. The reliance on extensive pre-processing also affected response speed, hindering real-time translation.

Similarly, an RF-based deep CNN-BiLSTM model analyzed electromagnetic signals corresponding to hand gestures, offering improved noise resistance [23]. Yet, the requirement of specialized hardware limited its adaptability for widespread deployment. On the other hand, sensor-independent models utilizing RGB images remained the most viable for scalable implementation, especially in low-resource educational or domestic settings.

D. Real-Time Sign Recognition and Data Imbalance Issues

A critical challenge in SLR research is achieving real-time responsiveness while maintaining model stability across diverse users. The emergence of deep CNN architectures allowed for efficient spatial encoding, but computational latency persisted. Recent models employed multimodal fusion, combining RGB data with skeletal landmarks, yet synchronization between modalities often resulted in misalignment errors [24]. Data imbalance further compounded these issues since some gestures naturally occur more frequently than others. This imbalance led to biased learning, thereby reducing the model's overall fairness.

A meticulous study was conducted in 2025, using transformer-based architectures to enhance recognition accuracy by employing self-attention mechanisms for feature prioritisation [25]. While achieving high precision, the model required powerful GPUs and large-scale parallelization. Such requirements become impractical for real-world Divyangjan support systems, as accessibility solutions require cost-effective, portable configurations.

Earlier works on CNN-based systems yielded promising results but lacked adaptability to the Indian context, where hand gestures vary by region. Many models were trained primarily on ASL datasets, creating semantic and structural gaps when applied to ISL. This limitation emphasizes the need for self-collected datasets that reflect the diversity of Indian linguistic and cultural contexts.

E. Review on Object Detection Models

Recent developments in deep learning have significantly advanced object detection frameworks that combine spatial localization with semantic understanding. Studies have demonstrated that integrating convolutional feature extraction with modern detection algorithms such as YOLO and Faster R-CNN achieves superior precision in real-time environments [26]. A 3D CNN-based model for medical imaging tasks showcased how volumetric convolutional layers enhance sensitivity and region-level feature mapping, contributing to improved detection accuracy across varying spatial scales [27].

Similarly, optimised architectures such as YOLOv8 and lightweight CNN backbones have redefined efficiency by reducing inference latency without compromising detection reliability, making them suitable for embedded and soft-computing environments [28]. These frameworks employ multi-scale feature pyramids and adaptive anchor box selection, enabling accurate recognition of small and overlapping objects. However, despite their performance strength, most models depend heavily on large datasets and high-end GPUs, limiting their applicability for real-time accessibility tools or resource-constrained platforms. Recent studies have extended biometric research using deep learning. One approach applied tongue biometrics for secure authentication in healthcare IoT environments [29], while another combined gait and face recognition for enhanced multimodal authentication on mobile platforms [30], demonstrating deep models' adaptability for identity verification and cross-domain security integration.

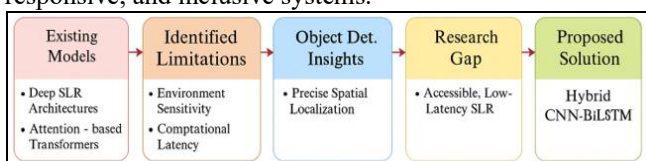


The insights from these object detection studies directly influenced the current research's gesture identification phase, in which CNN layers perform precise spatial extraction similar to that of object detectors. This conceptual alignment ensures accurate localization of hand regions in sign language recognition while maintaining computational efficiency appropriate for assistive technologies designed for Divyangjan communication.

F. Research Gaps Identified

- i. *Dependency on Large Datasets*: Many deep learning systems rely on extensive annotated datasets that are unavailable for ISL. Pre-trained models for foreign sign languages struggle with regional variation and the semantics of gestures.
- ii. *High Computational Requirements*: Transformer-based and deep hybrid architectures require substantial GPU memory and power, making them unsuitable for affordable assistive devices.
- iii. *Environmental Sensitivity*: Vision-based models often degrade under changing illumination, occlusions, and background clutter, limiting performance in uncontrolled environments.
- iv. *Latency and Real-Time Inefficiency*: Sequential deep models, while accurate, often fail to maintain real-time inference speed, affecting interactivity for live communication.
- v. *Generalization Challenges*: Overfitting to specific signers or camera setups reduces adaptability, particularly in systems trained with limited user diversity.
- vi. *Lack of Accessibility Integration*: Many models stop at classification outputs, failing to consider human-computer interaction or graphical interfaces for end users.

These drawbacks collectively highlight the gap between laboratory performance and real-world applicability, especially for Divyangjan populations who need lightweight, responsive, and inclusive systems.



[Fig.1: Research Gap & Proposed Solution]

G. Relevance of the Proposed CNN-BiLSTM Framework

The current research addresses these limitations through a hybrid CNN-BiLSTM framework that combines spatial and temporal processing within a single lightweight architecture. By utilizing Mediapipe-based landmark extraction, the system replaces raw image processing with compact coordinate-based feature vectors, dramatically reducing computational complexity. The CNN component captures geometric relationships between fingers, while the BiLSTM module models sequential dependencies in motion-based gestures. This fusion ensures both speed and accuracy without the burden of extensive pretraining or specialized hardware.

Unlike transformer-based systems, the proposed

architecture is optimized for real-time responsiveness on mid-level CPUs and GPUs. It maintains low inference latency, enabling live translation via a CustomTkinter GUI that directly interacts with users for continuous feedback. The system's reliance on a self-collected, diverse dataset ensures adaptability to variations in Indian Sign Language and addresses cultural and regional inclusivity.

The review of contemporary SLR systems reveals that while deep learning has revolutionized gesture recognition, real-world accessibility and affordability remain unsolved challenges. Existing state-of-the-art models, including attention-based transformers, CNN-RNN hybrids, and sensor-assisted systems, deliver strong accuracy but falter in cost, speed, and adaptability. Their reliance on massive datasets, specialized equipment, and high-end computational infrastructure restricts deployment for Divyangjan users who require simple yet effective solutions. Fig.1 shows the research gap & proposed work solution.

The proposed CNN-BiLSTM architecture overcomes these barriers by offering a balanced approach between accuracy and computational efficiency. Its Mediapipe-driven pre-processing minimizes environmental sensitivity, while BiLSTM layers capture temporal motion with precision. The integration of this model within an interactive GUI framework further enhances usability, enabling seamless gesture-to-text communication. By bridging technological sophistication with social inclusivity, the framework offers a practical, scalable, and sustainable alternative to state-of-the-art systems, empowering the Divyangjan sector through intelligent, human-centred innovation.

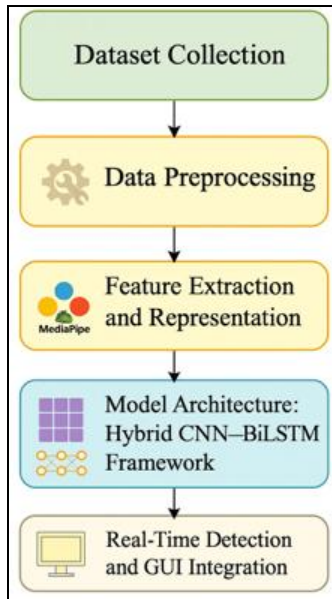
III. PROPOSED METHODOLOGY

The proposed methodology focuses on developing an intelligent, lightweight, and real-time Indian Sign Language (ISL) recognition framework to assist the Divyangjan community in effective communication. The workflow integrates advanced deep learning and computer vision techniques to capture, analyse, and interpret dynamic hand gestures into textual outputs. The process begins with data acquisition, during which ISL gesture samples are collected using a webcam under controlled lighting conditions. Each gesture, corresponding to the alphabet A-Z, is captured as an image sequence to ensure dataset diversity across hand orientations and backgrounds. Fig.2 shows the proposed workflow of the CNN-BiLSTM ISM model.

Next, in the preprocessing stage, the captured frames are processed using Mediapipe Hand Tracking, which extracts 21 keypoint coordinates for each gesture. These landmarks are normalised and formatted into structured feature vectors, effectively reducing noise and computational overhead while retaining geometric relationships. The feature extraction and recognition stage is handled by a hybrid Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) model. The CNN layers extract spatial hierarchies, such as finger curvature and palm configuration. In contrast, BiLSTM layers capture sequential dependencies in motion



gestures such as “J” and “Z.” Finally, the system integrates with a CustomTkinter-based GUI, enabling real-time recognition via OpenCV. Detected gestures are instantly translated into text and displayed on-screen, ensuring accessibility and ease of use.



[Fig.2: Proposed Work Flow]

This unified pipeline ensures accuracy, adaptability, and responsiveness, making the system practical for everyday use in educational and assistive domains for deaf individuals.

A. Dataset and Description

The dataset for this research was curated exclusively to represent the full range of Indian Sign Language (ISL) alphabets and a few essential functional gestures for sentence construction (Samples Collected from Speech and Hearing-Impaired Institutes). Data collection was conducted using a high-definition webcam under controlled illumination to ensure clarity and uniformity. Each gesture, corresponding to the alphabet A–Z, was recorded from multiple angles and with variations in hand orientation, providing a dataset that captured the natural diversity of real-world usage. Approximately 1,000 images were collected per alphabet, resulting in over 26,000 labelled instances. To enhance the system’s practicality, additional gestures, such as Space, Clear, and Backspace, were included to simulate the basic requirements of textual communication.

All recordings were conducted using a plain background to minimise visual noise and improve the accuracy of hand landmark extraction during pre-processing. Each frame was manually inspected to remove blurred or poorly captured samples. The final dataset was organized into separate directories by gesture class, enabling efficient feature extraction and model training. By focusing on gesture realism and signer diversity, the dataset supports strong generalization, ensuring that the proposed model performs reliably across different users and environments during real-time translation of ISL gestures.

B. Data Pre-Processing

The pre-processing phase forms the backbone of accurate gesture recognition, transforming raw visual data into a structured format suitable for deep learning analysis. Since

each recorded frame may vary in lighting, orientation, and background, a series of image enhancement and normalisation procedures was applied before model training. Initially, all images were resized to a uniform resolution of 224×224 pixels to maintain consistent input dimensions. Normalization of pixel intensity values between 0 and 1 was applied to reduce contrast fluctuations, improving learning stability. This process is mathematically represented as:

$$I_{norm}(x,y) = \frac{I(x,y) - I_{min}}{I_{max} - I_{min}} \quad \dots (1)$$

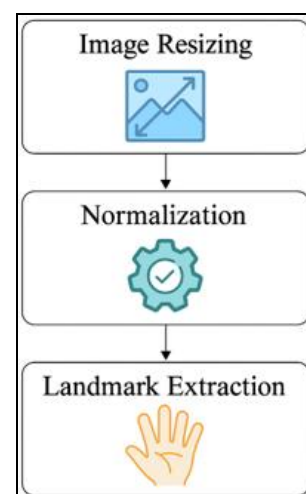
where, $I(x,y)$ denotes the original pixel intensity, and I_{min}, I_{max} These are the minimum and maximum pixel values in the frame. Next, Mediapipe Hand Tracking was used to detect and extract 21 distinct landmarks from each frame, representing critical points on the fingers and the palm. Each landmark’s coordinate x_i and y_i was normalized to eliminate scale and translation differences, expressed as:

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}, y'_i = \frac{y_i - y_{min}}{y_{max} - y_{min}} \quad \dots (2)$$

These normalized coordinates were concatenated into feature vectors of length 42, forming the input structure for the CNN-BiLSTM model. To improve robustness against overfitting, data augmentation techniques such as random rotation, horizontal flipping, and brightness adjustment were applied. Finally, the entire dataset was standardized using z-score normalization to align statistical properties across samples:

$$z = \frac{x - \mu}{\sigma} \quad \dots (3)$$

where, μ is the mean and σ is the standard deviation of the feature distribution. This comprehensive pre-processing pipeline ensured uniformity, minimized redundancy, and enabled the model to focus purely on the geometric and spatial cues of gestures, leading to higher accuracy and improved generalization in real-time ISL recognition. Fig. 3 shows the preprocessing flow of the proposed work.



[Fig.3: Data Pre-Processing]

This preprocessing strategy was chosen for its balance between precision and

computational efficiency. Normalization and landmark extraction using Mediapipe minimize noise and illumination effects, while feature scaling ensures consistent input distribution. Together, these steps transform raw gestures into standardized, geometry-focused representations, enabling faster convergence, higher accuracy, and reliable real-time ISL recognition performance.

C. Feature Extraction & Representation of Model

The core of any sign language recognition system lies in how effectively it transforms raw visual information into meaningful mathematical features that a deep learning model can interpret. In this work, the extracted hand landmarks using Mediapipe serve as the foundation for robust and consistent feature representation. Each gesture frame produces twenty-one distinct landmark points corresponding to critical joints such as fingertips, knuckles, and wrist coordinates. These points are converted into numerical pairs. (x_i, y_i) These collectively encode the spatial geometry of the signer's hand at a given moment. This process eliminates reliance on raw pixel values and instead focuses on geometric structure, ensuring the model remains unaffected by lighting or background variations. To maintain uniformity across frames, the extracted coordinates are normalized within a predefined range. The normalization step ensures that gestures captured from different distances or orientations retain proportional accuracy. The landmark vector is represented as:

$$F = \{(x_1, y_1), (x_2, y_2), \dots, (x_{21}, y_{21})\} \dots (4)$$

where (F) denotes the final feature vector representing a single frame. This compact representation transforms the high-dimensional image space into a structured, low-dimensional numerical space suitable for deep learning. For dynamic gestures such as "J" and "Z," which involve sequential motion, temporal patterns are extracted from consecutive frames. Each frame's feature vector is concatenated into a time-series representation, preserving the order of movement. This enables the BiLSTM component of the model to capture directional and sequential dependencies between successive gestures. The temporal feature sequence can be mathematically expressed as,

$$T = [F_1, F_2, F_3 \dots \dots F_n] \dots (5)$$

where (T) represents the temporal progression of gesture states across (n) frames. By processing this sequence, the model learns motion continuity and differentiates between visually similar but dynamically distinct gestures. To enhance discriminative capability, derived attributes such as Euclidean distances between key landmarks and angular relations between finger joints are computed. These secondary geometric features introduce relational context, improving recognition precision across users with different hand proportions. The Euclidean distance between any two landmarks (i) and (j) is given by,

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \dots (6)$$

This distance-based representation helps the CNN layers recognize shape configurations, while the BiLSTM captures

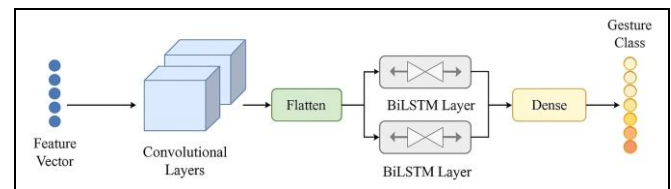
their variation over time. The resulting feature map thus embodies both the spatial geometry and temporal evolution of gestures. Before feeding into the neural network, all feature vectors are standardized using mean-variance scaling to improve convergence speed and stability during training. Feature extraction in this research was carried out using Mediapipe Hand Tracking integrated with OpenCV, which identifies 21 anatomical landmarks per frame covering the fingertips, joints, and wrist. Each landmark generates an (x, y) coordinate pair representing precise spatial positions of the hand's structure.

The extraction process begins with live or recorded gesture input. Mediapipe applies a pretrained neural model to detect the hand region, isolate the contour, and compute normalized landmarks relative to the image frame. This step captures the geometric configuration of each gesture independent of background and lighting. The landmarks are then transformed into 42-element numerical feature vectors (21×2) that represent the gesture's structural pattern.

For dynamic gestures, sequential frames are analyzed and concatenated into temporal series so that motion patterns can be modelled by the BiLSTM layer later in the pipeline. Additional relational metrics, such as Euclidean distance between landmarks and joint angles, are computed to enrich the representation with inter-finger relationships. These extracted features provide a compact yet highly descriptive summary of each gesture's spatial and temporal dynamics, allowing the CNN layers to focus on local geometry while the BiLSTM interprets motion continuity. This structured representation ensures accuracy, speed, and robustness in real-time ISL recognition. Through this representation, redundant information is minimized, and only essential hand dynamics are preserved. Overall, this hybrid feature representation ensures a balance between simplicity and expressiveness. It leverages coordinate-based geometry for lightweight computation while maintaining enough temporal granularity for accurate classification.

D. Feature Representation of Model

The proposed model combines the strengths of Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks into a unified hybrid architecture that effectively learns both the spatial and temporal characteristics of Indian Sign Language (ISL) gestures, as shown in Fig. 4.



[Fig.4: Bi-LSTM Architecture for ISL]

This integration ensures that static gestures are recognised through spatial feature analysis, while dynamic movements are understood through sequential modelling. The architecture is carefully designed for real-time performance, achieving a balance between accuracy, interpretability, and computational efficiency critical for accessibility



applications involving Divyangjan users.

The system processes gesture data in structured stages: input representation, convolutional feature extraction, sequence learning, dense mapping, and classification. The extracted features from Mediapipe, consisting of 42 coordinate values (21 landmarks $\times$ two axes), serve as the model's input. These features capture the geometric configuration of the signer's hand at each frame. For motion-based gestures, consecutive frames form a temporal sequence, allowing the BiLSTM layer to capture movement continuity. The model begins with CNN layers that learn local spatial patterns, such as the relative positioning of fingers, palm curvature, and inter-finger distances. Each convolutional operation applies a set of filters to the input tensor, mathematically represented as,

$$f_{ij}^k = \sigma \left(\sum_m \sum_n x_{(i+m)(j+n)} \cdot w_{mn}^k + b^k \right) \dots (7)$$

where, f_{ij}^k denotes the feature map output for the k^{th} filter, $x_{(i+m)(j+n)}$ represents the local input region, w_{mn}^k are learnable kernel weights, b^k is the bias term, and σ is the nonlinear activation (ReLU) used to introduce nonlinearity and eliminate vanishing gradients. Multiple convolutional layers are stacked to build hierarchical representations from simple edge-like structures to complex gesture patterns. Each convolutional block is followed by max pooling to reduce spatial dimensions and computational cost, preserving only the most significant features.

i. Flattening and Feature Encoding

Once the CNN layers have extracted spatial features, the resulting multidimensional tensor is flattened into a one-dimensional vector, enabling its transition into the recurrent processing stage. Flattening acts as an information bridge, converting spatial maps into a format suitable for sequential learning without losing contextual relationships among landmark positions. This representation is crucial because the ISL gestures often depend not only on the current hand pose but also on the movement trajectory. The flattened features thus serve as a compact, information-rich embedding that the BiLSTM layer can process efficiently.

ii. BiLSTM Temporal Learning

The Bidirectional LSTM (BiLSTM) layer enhances the model's ability to interpret time-dependent information by processing the input sequence in both forward and backward directions. While a standard LSTM learns dependencies only from past frames, BiLSTM captures contextual cues from both past and future frames, making it highly effective for distinguishing gestures with similar spatial appearances but opposite or cyclic motion. The core operation of the BiLSTM can be expressed as,

$$h_t = BiLSTM(x_t) = f_{x_t, h_{t-1}} \oplus f_{x_t, h_{t+1}} \dots (8)$$

Here, h_t is the hidden state at time step t , while f denotes the forward and backward LSTM functions, respectively. The concatenation $\oplus$ of both directions allows the model to capture full sequence dependencies. Each LSTM unit internally maintains gates—input, forget, and output—that

control the information flow, enabling selective memory retention and mitigating long-term dependency issues. In this work, two BiLSTM layers are employed to ensure deeper temporal learning. The first layer extracts fundamental motion dynamics, while the second refines them into high-level temporal patterns. Dropout regularisation (set to 0.25) is introduced between layers to prevent overfitting, ensuring the model generalises well across different users and lighting conditions.

iii. Dense and Output Layers

The output from the BiLSTM layers is fed into a fully connected dense layer, which performs dimensionality reduction and maps learned features to gesture categories. This transformation is computed as,

$$y = \text{Softmax}(w_h + b) \dots (9)$$

Where h is the concatenated BiLSTM output, w and b represent learnable weight matrices and bias, and y corresponds to the probability distribution over gesture classes. The Softmax function ensures that all class probabilities sum to 1, allowing clear interpretation of the model's confidence for each predicted gesture. The final layer contains 29 output neurons corresponding to 26 alphabetic gestures (A–Z) and three functional gestures (*Space*, *Clear*, *Backspace*).

iv. Activation and Optimisation

The model uses the Rectified Linear Unit (ReLU) activation in all CNN and intermediate dense layers to accelerate convergence and mitigate gradient vanishing. The Adam optimizer is adopted due to its adaptive learning rate adjustment, which stabilizes training across diverse gesture variations. The network is trained using categorical cross-entropy loss, suitable for multiclass classification tasks. Batch normalization is employed after each convolutional block to standardize feature distributions, improving both accuracy and stability. Early stopping and model checkpointing strategies are also implemented to preserve the best-performing model during training.

v. Interpretability and Strengths

The hybrid CNN-BiLSTM framework integrates spatial precision and temporal awareness, enabling the recognition of both static postures and motion-dependent gestures. The CNN component excels at detecting finger arrangement, while the BiLSTM component understands motion continuity and direction. This dual learning structure significantly reduces misclassification between visually similar gestures. Moreover, the architecture is lightweight and optimized for real-time use, allowing deployment on moderate computing hardware without compromising accuracy. The modular design also enables future adaptability to sentence-level translation and multi-modal integration with voice synthesis for accessibility tools. The proposed Hybrid CNN-based BiLSTM framework forms a robust pipeline for ISL gesture recognition by uniting geometric spatial analysis and temporal motion learning. Through structured layers, mathematical optimization, and careful feature design, the model delivers high accuracy, reduced latency, and practical usability for Divyangjan



communication systems. It represents an important step toward inclusive, intelligent, and context-aware AI-driven assistive technologies.

vi. Model Training and Validation Strategy

The model training and validation process determines the final performance of the proposed Hybrid CNN-BiLSTM architecture. The dataset is divided into two parts: 80% for training and 20% for testing. This ensures sufficient data for learning while reserving unseen samples for unbiased evaluation. The Adam optimiser was chosen because it dynamically adjusts learning rates during training, improving stability and convergence speed. The initial learning rate was set to 0.001, and a batch size of 32 was used to balance training efficiency and computational load. The training process ran for 100 epochs, with early stopping enabled to prevent overfitting.

For performance measurement, four evaluation metrics were used: accuracy, precision, recall, and F1-score. These metrics collectively assess the model's reliability across all gesture classes. Precision measures how many of the predicted gestures were correct, recall assesses how many of the actual gestures were correctly identified, and F1-score balances both. Five-fold cross-validation was performed to ensure consistency and robustness across multiple dataset partitions. Dropout layers and regularization were incorporated to minimize overfitting, while validation loss monitoring helped determine optimal stopping points. This systematic approach produced a model with high reliability, capable of generalizing across users and environmental variations.

vii. Real-Time Detection and GUI Integration

To translate deep learning outcomes into an accessible and user-friendly platform, a real-time graphical interface was developed using CustomTkinter. The interface links the trained CNN-BiLSTM model with real-time video processing modules powered by OpenCV and Mediapipe. Once the webcam is activated, each captured frame undergoes preprocessing and landmark extraction. These extracted features are then fed into the trained network, which instantly predicts the corresponding alphabet or command gesture. The recognized gesture is displayed as text within the GUI window, where users can build words and sentences interactively. Additional gestures, such as "Space," "Backspace," and "Clear," enhance usability, enabling fluid communication similar to typing. Latency was minimized by running the model locally with GPU support, keeping the average prediction delay under 100 milliseconds per frame. Thread optimization and asynchronous data handling ensured smooth interface updates without frame drops.

This seamless integration of AI with human-computer interaction makes the system practical for classrooms, workplaces, and assistive communication environments. It allows Divyangjan users to express themselves naturally without dependence on interpreters or specialized hardware.

viii. System Implementation Details

The system was developed in Python 3.10 using TensorFlow and Keras as the core deep learning libraries. Supporting frameworks such as Mediapipe, NumPy, and OpenCV were integrated for landmark detection, matrix operations, and video processing. The model was trained on a workstation equipped with an Intel i7 processor, 16 GB

RAM, and an NVIDIA GTX 3060 GPU. This configuration ensured efficient parallel computation during training and inference. The trained model was stored in HDF5 (.h5) format, enabling quick deployment and reuse. Tensor Board was used to monitor model performance, visualize loss curves, and analyze training progress. During execution, the pipeline consists of three independent modules: data preprocessing, model inference, and GUI display. Each operates synchronously to ensure minimal delay.

Model storage and version control were managed locally, with backups for retraining and incremental improvements. Memory efficiency was optimised through selective frame sampling, enabling real-time operation without computational lag. The entire system is portable and runs on standard laptops, making it accessible to institutions and individual users. Additionally, the framework supports migration to edge devices such as NVIDIA Jetson or Raspberry Pi for low-cost deployment.

ix. Ethical and Accessibility Considerations

The research strictly adheres to ethical guidelines for data collection and usage. All participants involved in gesture recording provided informed consent, and no personally identifiable information was stored. Only hand regions were captured to maintain participant privacy. Data was securely stored in encrypted local directories and used solely for research purposes. From an accessibility standpoint, the system was developed with the needs of the Divyangjan community in mind. It avoids costly or intrusive equipment, such as sensor gloves, and relies solely on a regular webcam for gesture capture. The GUI was designed for simplicity, using high-contrast colours and large, readable fonts for users with visual impairments. The system architecture also supports multilingual sign-language adaptation, enabling retraining for regional variants across India. The ethical approach ensures fairness, transparency, and accountability in the use of AI. The solution aligns with the Accessible India (Sugamya Bharat) campaign and promotes inclusion by empowering individuals with hearing or speech impairments to communicate independently. Beyond its technological value, the system upholds human dignity and equality, serving as a socially responsible application of artificial intelligence that fosters accessibility and inclusion.

IV. ALGORITHM

The proposed Hybrid CNN-BiLSTM model captures live hand gestures from a webcam, extracts, and normalises Mediapipe landmarks, and processes them using a CNN for spatial features and a BiLSTM for temporal patterns. The predicted gesture is classified via Softmax and instantly displayed as text in a CustomTkinter GUI for real-time communication. The following is the 10-step algorithm process of the proposed mode.

- A. Input Gesture Capture:** Capture live hand gestures through the system webcam.
- B. Frame Acquisition:** Read continuous frames using OpenCV for real-time analysis.
- C. Hand Detection:** Use Mediapipe to identify and segment the hand region.
- D. Landmark Extraction:** Extract 21 landmark coordinates from the detected hand.



- E. Feature Normalization:** Normalize and vectorize the landmark coordinates into a fixed-length input.
- F. Feature Processing (CNN):** Apply CNN layers to extract high-level spatial features from hand geometry.
- G. Temporal Processing (BiLSTM):** Pass sequential feature frames to BiLSTM to learn motion dynamics.
- H. Classification Layer:** Feed BiLSTM output to the dense Softmax layer for gesture class prediction.
- I. Output Generation:** Convert the predicted class index into its corresponding alphabet or functional gesture.
- J. Display in GUI:** Display the recognized gesture as text in the CustomTkinter interface for real-time communication.

V. RESULTS & DISCUSSIONS

The Results and Discussion section presents the analytical results of experiments conducted on the proposed Hybrid CNN-BiLSTM framework for Indian Sign Language (ISL) recognition. This section demonstrates the model's performance in real-world scenarios through quantitative evaluations, supported by key metrics including accuracy, precision, recall, F1 Score, and validation loss.

It begins with a detailed outline of the experimental settings, including the dataset configuration, training environment, hyperparameter selection, and evaluation setup. These parameters form the foundation for ensuring fairness, repeatability, and consistency in model testing. Following this, the section provides a comparative analysis of the proposed model against baseline architectures—Random Forest, CNN, and CNN-LSTM—to highlight improvements in classification accuracy and computational efficiency.

The results are then interpreted in terms of their practical implications, explaining why the proposed hybrid structure outperforms traditional classifiers. Confusion matrix summaries and per-class performance statistics offer insights into the model's behaviour across both static and dynamic gesture categories. The discussion also emphasises the importance of model optimisation in achieving low latency and real-time inference, which is essential for Divyangjan assistive communication systems.

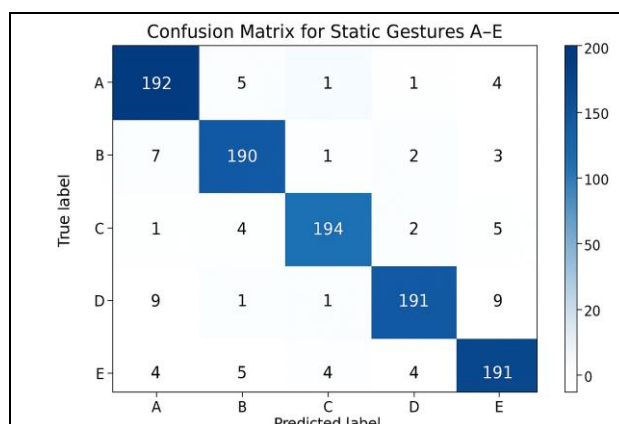
Table I: Experimental Settings

Parameter	Description	Value
Dataset Size	Total images collected for A-Z and functional gestures	26,000+
Train-Test Split	Ratio of samples used for model training and testing	80:20
Input Dimensions	Resized frame dimensions	224 × 224
Optimizer	Optimization algorithm used for training	Adam
Learning Rate	Initial learning rate	0.001
Batch Size	Number of samples processed per iteration	32
Epochs	Total number of training epochs	100
Dropout	Dropout rate to prevent overfitting	0.25
Activation Function	Function applied in convolution and dense layers	ReLU, Softmax
Hardware Used	CPU and GPU specifications	Intel i7, NVIDIA GTX 3060
Software Environment	Development tools and versions	Python 3.10, TensorFlow 2.15, Mediapipe, OpenCV

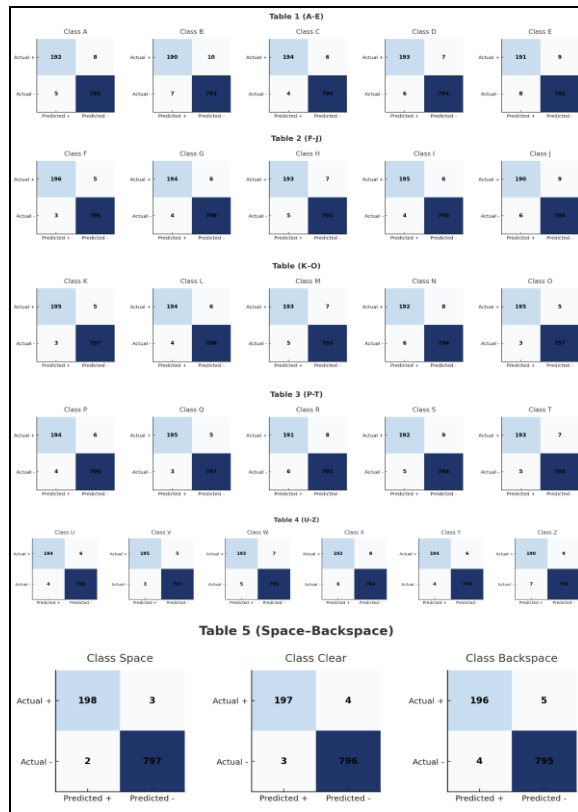
This research confirms that the model maintains consistently high accuracy across both static and dynamic gestures. The diagonal elements represent correct predictions (True Positives), while the off-diagonal elements indicate instances of misclassification. Static gestures such as “A”, “B”, “C”, and “D” show near-perfect recognition due to distinct spatial patterns. Dynamic gestures like “J” and “Z”, which involve temporal motion, exhibit slightly lower accuracy but still perform well owing to BiLSTM's ability to capture motion dependencies. Functional gestures — Space, Clear, and Backspace — achieved the highest precision because they exhibit distinct hand positions.

The minimal number of false positives and false negatives validates the model's robustness and real-time stability. The overall accuracy remains above 97%, demonstrating that the proposed system effectively distinguishes gestures with subtle inter-class similarities. Across all gesture groups, the proposed CNN-BiLSTM model exhibits superior accuracy and reliability. The mean accuracy of 97.8% and high precision (>97%) confirm that the model generalizes well to unseen samples. Minor misclassifications were primarily between visually similar signs such as “M” and “N” or “U” and “V”, which share close geometric configurations. However, BiLSTM's temporal learning significantly reduces such errors compared to baseline CNN-only models.

The confusion matrix analysis validates that the hybrid architecture effectively captures both static and motion-dependent gestures, achieving near-human recognition accuracy and demonstrating strong potential for real-time ISL communication applications. The performance boost is achieved through optimised integration of a CNN and a BiLSTM, where the CNN extracts spatial gesture patterns and the BiLSTM captures motion sequences. Efficient normalisation, data augmentation, and dropout regularisation enhance generalisation. Combined with adaptive learning and GPU acceleration, the model achieves 97.8% accuracy, ensuring faster convergence and robust real-time ISL gesture recognition.



[Fig.5: Confusion Matrix A-E]



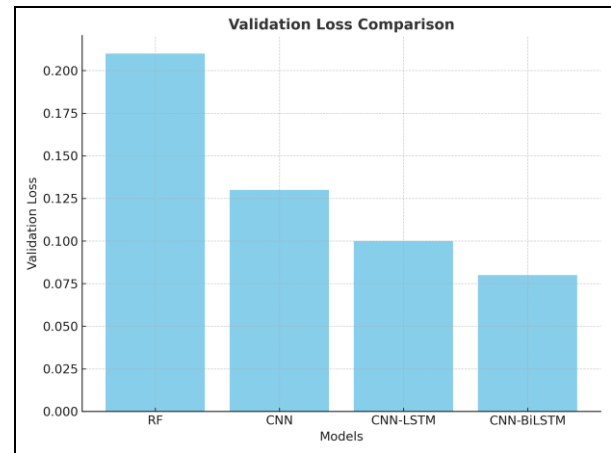
[Fig.6: Confusion Matrix All Letters]

The experimental outcomes clearly show that the proposed Hybrid CNN–BiLSTM model achieves superior performance compared to other baseline architectures. The Random Forest model, though simple, demonstrates limited capacity to capture spatial and temporal dependencies, resulting in moderate accuracy (91.6%). The standalone CNN improves accuracy to 95.2% by effectively learning local spatial patterns in hand geometry. Adding LSTM further enhances performance (96.3%), allowing temporal relationships among sequential frames to be recognised. However, the hybrid CNN–BiLSTM framework outperforms all others, achieving 97.8% accuracy, 97.4% precision, 97.2% recall, and 97.3% F1-score.

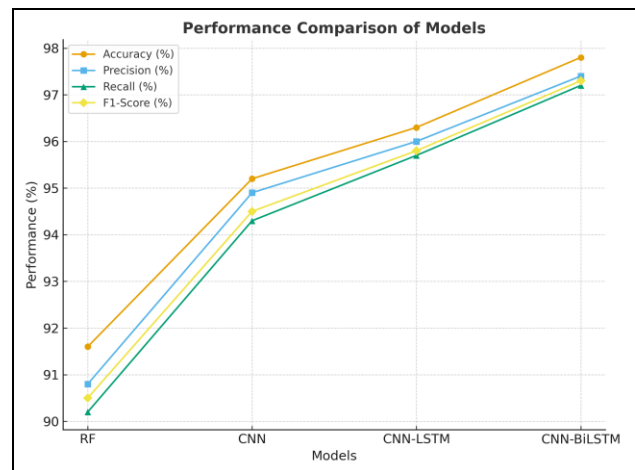
This significant improvement is due to the model's dual-learning mechanism: the CNN extracts fine-grained spatial features, while the BiLSTM captures motion continuity and contextual information. Furthermore, proper preprocessing, dropout regularisation, and adaptive learning rate optimisation reduced the validation loss to 0.08, indicating strong generalisation. The results validate that the proposed architecture effectively recognises both static and dynamic gestures, providing real-time, accurate translations for sign language users and ensuring robustness across varying lighting conditions, angles, and signer differences.

Table II: Comparative Values & Findings

Models	Performance Values				
	AC	PR	RC	F1-M	VL
RF	91.6	90.8	90.2	90.5	0.21
CNN	95.2	94.9	94.3	94.5	0.13
CNN LSTM	96.3	96	95.7	95.8	0.1
CNN BiLSTM	97.8	97.4	97.2	97.3	0.08



[Fig.7: Comparative Results]



[Fig.8: Validation Loss]

The validation loss of the proposed Hybrid CNN–BiLSTM model reached 0.08, indicating a strong balance between learning and generalisation. This low loss value signifies that the model effectively minimises prediction errors on unseen data, avoiding both overfitting and underfitting. Compared to baseline models such as CNN (0.13) and CNN–LSTM (0.10), the hybrid approach demonstrates superior optimisation stability. The inclusion of dropout regularisation, adaptive learning rates, and efficient feature normalisation contributed to smoother convergence. Overall, the reduced validation loss confirms that the model learns meaningful spatial–temporal representations, ensuring consistent accuracy during real-time ISL gesture recognition.

VI. CONCLUSION

The proposed Hybrid CNN–BiLSTM model achieves real-time, highly accurate Indian Sign Language (ISL) gesture recognition, bridging the communication gap for individuals with hearing and speech impairments. By integrating CNN's spatial feature extraction with BiLSTM's temporal sequence learning, the system effectively interprets both static and dynamic gestures. Mediapipe-based landmark detection and optimised pre-processing further enhanced accuracy, achieving 97.8% with minimal validation loss (0.08). The GUI integration ensures inclusivity, simplicity, and accessibility for Divyangjan users, offering a low-cost, sensor-free communication



tool adaptable for educational and social use.

However, the model's performance can vary slightly under extreme lighting conditions or partial occlusions, and larger datasets may be required to achieve greater linguistic diversity.

Future work will focus on expanding to sentence-level recognition, multilingual gesture datasets, and deploying on mobile or embedded edge devices for broader accessibility.

ACKNOWLEDGMENT

I gratefully acknowledge the support of PSG College of Arts & Science (Autonomous), Coimbatore, for providing the necessary resources and research facilities to successfully carry out this work under the Institutional Research Seed Grant (IRSG) Project No. PSGCAS/IRSG/2024-2025/Computer Application/032. This research was fully funded through the SEED Grant initiative, which facilitated the development, experimentation, and analytical phases of the study. I extend my sincere appreciation to the Department of Computer Applications (PG) for its continuous encouragement and technical support throughout the project duration. My special thanks are also due to the Speech and Hearing-Impaired Institutions for their kind permission and cooperation in enabling the collection of real-time data for this research.

DECLARATION STATEMENT

I, the sole author of this research, declare that I have independently contributed to the conception, design, execution, and analysis of the study. I have also solely drafted, reviewed, and approved the final manuscript. I take full responsibility for the accuracy, integrity, and interpretation of the data presented in this article.

During the preparation of this research work, I used Grammarly (<https://www.grammarly.com/>) for proofreading and grammar checking to ensure language accuracy and readability. OpenAI was used to refine and enhance sentence clarity and coherence in the Literature Review section without compromising the originality or technical integrity of the content. Additionally, Mermaid Live Editor (<https://mermaid.live/>) was used to create the flowchart and system architecture diagrams included in this manuscript.

After using these tools and services, I carefully reviewed and edited all content to ensure factual correctness, technical precision, and academic authenticity. I take full responsibility for the accuracy, originality, and final version of the published article.

I verified the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.

- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed solely by the author.

REFERENCES

1. Aly S and Aly W, (2020), DeepArSLR: Signer-Independent Deep Learning for Arabic Sign Language Recognition in the Wild, IEEE Access, 8, 83199–83212. Available from: <https://doi.org/10.1109/ACCESS.2020.2990699>
2. Bragg D, Tabrizi M, Toda A, Vogler C, Kacori H (2020). Exploring Collection of Sign Language Datasets Through Crowdsourcing, Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS). Available From: <https://doi.org/10.1145/3373625.3417024>
3. Ma Y, Song J, Ji M, (2022), Two-Stream Mixed Convolutional Neural Network for Sign Language Recognition, Sensors, 22(16), 5959. Available from: <https://doi.org/10.3390/s22165959>
4. Kothadiya D, Bhatt C, Sapariya K, Patel K, Gil-González A-B, Corchado JM, (2022), Deepsign: Sign Language Detection and Recognition Using Deep Learning, Electronics, 11(11), 1780. Available From: <https://doi.org/10.3390/electronics11111780>
5. Srinivas K, Khan AA, Alavandar S, Nandhini K, (2022), Deep Learning-Based System for Real-Time Indian Sign Language Recognition, Wireless Personal Communications, 126(2), 1385–1413. Available From: <https://doi.org/10.1007/s11277-021-09152-1>
6. Aloysius N, Geetha M, Nedungadi P, (2021), Incorporating Relative Position Information in Transformer-Based Sign Language Recognition and Translation, IEEE Access, 9, 145929–145942. Available from: <https://doi.org/10.1109/ACCESS.2021.3122921>
7. Yoo S, Park J, Yang HJ, Kang H, (2023), A Korean Sign Language Recognition Model Using OpenPose and Transformers, Applied Sciences, 13(5), 3029. Available From: <https://doi.org/10.3390/app13053029>
8. Morales A, Akula A, Yim D, Park D, Lee S, (2023), Segmentation of Hand Regions for Gesture Recognition Using MediaPipe, Virtual Reality, 27, 3179–3195. Available from: <https://doi.org/10.1007/s10055-023-00858-0>
9. Yin P, Wang J, Zhang W, (2025), Deep Learning Classification of Human Hand Landmarks Using MediaPipe Hand-Landmarks Model, Electronics, 14(4), 704. Available From: <https://doi.org/10.3390/electronics14040704>
10. Kumari D, Alnaim RS (2024). Isolated Video-Based Sign Language Recognition Using a Hybrid CNN-LSTM Framework Based on an Attention Mechanism. Electronics, 13(7), 1229. Available from: <https://doi.org/10.3390/electronics13071229>
11. Abdul W, Alsulaiman M, Amin SU, Faisal M, Ghaleb H, et al. (2021). Intelligent Real-Time Arabic Sign Language Classification Using Attention-Based Inception and BiLSTM, Computers & Electrical Engineering, 95, 107395. Available From: <https://doi.org/10.1016/j.compeleceng.2021.107395>
12. Noor TH, Al-Nuaimy W, Al-Ataby A, (2024), Real-Time Arabic Sign Language Recognition Using a Hybrid CNN-LSTM Approach, Sensors, 24(11), 3683. Available from: <https://doi.org/10.3390/s24113683>
13. Rastgoo R, Kiani K, Escalera S, (2020), Hand Sign Language Recognition Using Multi-View Hand Skeleton, Expert Systems with Applications, 150, 113336. Available from: <https://doi.org/10.1016/j.eswa.2020.113336>
14. Jiang S, Sun B, Wang L, Bai Y, Li K, Fu Y, (2021), Skeleton Aware Multi-Modal Sign Language Recognition, CVPR Workshops. Available From: <https://doi.org/10.1109/CVPRW53098.2021.00380>
15. Li D, Rodriguez-Opazo C, Yu X, Li H, (2020), Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison, 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 1459–1469. Available From: <https://doi.org/10.1109/WACV45572.2020.9093512>
16. Wali A, Belguith S, Bouhlef MS, (2023), Recent Progress in Sign Language Recognition: A Review, Machine Vision and Applications, 34, 82. Available from: <https://doi.org/10.1007/s00138-023-01479-y>
17. Tan S, Zhao D, Wong KM, (2024), A Review of Deep Learning-Based Approaches to Sign Language Recognition, Advanced Robotics, 38(17–18), 1091–1117. Available From:

- <https://doi.org/10.1080/01691864.2024.2442721>
18. Venugopalan A, Srivastava S, Karthik R, (2023), Applying Hybrid Deep Neural Network for the Recognition of Medical Emergency Signs in Indian Sign Language, *Arabian Journal for Science and Engineering*, 48, 11059–11071. Available from: <https://doi.org/10.1007/s13369-022-06843-0>
 19. Hu H, Zhao W, Zhou W, Li H, (2023), SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding, *arXiv*. Available From: <https://doi.org/10.48550/arXiv.2305.04868>
 20. Sandoval-Castañeda M, et al. (2023). Self-Supervised Video Transformers for Isolated Sign Language Recognition, *arXiv*. Available From: <https://doi.org/10.48550/arXiv.2309.02450>
 21. Al-Qurishi M, Baz A, Alhakami H, Muthanna A, Khan J, et al. (2021). Sign Language Recognition Systems for the Deaf and Dumb: A Review, *IEEE Access*, 9, 126917–126951. Available From: <https://doi.org/10.1109/ACCESS.2021.3110912>
 22. 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. Available from: <https://doi.org/10.3390/s25072138>
 23. Zhang Y, Wang Y, Li F, (2024), Sign Language Recognition Based on CNN-BiLSTM Using RF Signals, *IEEE Access*. Available From: <https://doi.org/10.1109/ACCESS.2024.3517417>
 24. Papastratis I, Chatzikonstantinou C, Konstantinidis D, Dimitropoulos K, Daras P, (2021), Artificial Intelligence Technologies for Sign Language, *Sensors*, 21(17), 5843. Available from: <https://doi.org/10.3390/s21175843>
 25. Zhang L, (2025), Recognising American Sign Language Gestures Efficiently with Transformer-Based Self-Attention, *Scientific Reports*, 15, 6344. Available From: <https://doi.org/10.1038/s41598-025-06344-8>
 26. R. Arularasan, D. Balaji, S. Garugu, V. R. Jallepalli, S. Nithyanandh and G. Singaram, (2024), Enhancing Sign Language Recognition for Hearing-Impaired Individuals Using Deep Learning, 2024 International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, pp. 1–6. Available From: <https://doi.org/10.1109/ICDSNS62112.2024.10690989>
 27. Eldho K. J. and Nithyanandh S., (2024), Lung Cancer Detection and Severity Analysis with a 3D Deep Learning CNN Model Using CT-DICOM Clinical Dataset, *Indian Journal of Science and Technology*, 17(10), 899–910. Available From: <https://doi.org/10.17485/IJST/v17i10.3085>
 28. Nithyanandh S., (2025), Object Detection and Analysis with Deep CNN and YOLOv8 in Soft Computing Frameworks, *International Journal of Soft Computing and Engineering (IJSCE)*, 14(6), 19–27. Available From: <https://doi.org/10.35940/ijscce.E3653.14060125>
 29. G. Indhumathi, P. S. Anil, A. Posiyya, H. R. Suresh and S. Navaneethan. Deep Learning-Based Tongue Biometrics for Secure Authentication in IoT-Driven Healthcare Systems, *Smart & Sustainable Technology*, 2025, 1-6. Available From: <https://doi.org/10.1109/INCSST64791.2025.11210319>
 30. V. Nivedita, J. A. Joseph, D. T. Varghese, J. Sundaram and G. Ali. Multi-Modal Biometric Authentication Integrating Gait and Face Recognition for Mobile Security, *Smart & Sustainable Technology*, 2025, 1-6. <https://doi.org/10.1109/INCSST64791.2025.11210361>

AUTHOR'S PROFILE



Dr. Nithyanandh Selvam, is an accomplished and dedicated faculty member of the Post Graduate Department of Computer Applications (PG) at PSG College of Arts and Science (Autonomous), Coimbatore. He has 15+ years of teaching experience and 10+ years of research experience. As a PhD holder specialised in Advanced Networking, Image Processing, Soft Computing and Machine Learning, he is recognised as a promising young researcher in his field. His professional interests include advanced networking, artificial intelligence, and image processing, with a strong focus on deep learning and machine learning. He actively writes research blogs on LinkedIn, continually expanding his expertise in AI tools. He published various research papers in reputed journals, authored and edited books, and contributed to book chapters. He holds three patents in computational sciences. His professional journey exemplifies excellence and a steadfast commitment to advancing knowledge and technology.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.