International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

SignSense AI: Transformer Based Indian Sign Language Recognition Using Key-Point Features

Dhulipala Koushik
Under Graduate CSE(AI&ML)
GCET
Hyderabad, India
22r11a66g2@gcet.edu.in

Pagadala Rahul Naidu
Under Graduate CSE(AI&ML)
GCET
Hyderabad, India
22r11a66h9@gcet.edu.in

Ms. Jeevitha Joseph
(Asst. Professor)
GCET
Hyderabad, India
jeevitha.cse@gcet.edu.in

ABSTRACT

Deaf and hard-of-hearing communities rely on sign languages as their primary mode of communication, yet automated systems capable of interpreting these languages accurately and efficiently remain limited — particularly for Indian Sign Language. This work presents a gesture recognition architecture that departs from conventional image-based pipelines by operating entirely on structured body and hand landmark sequences extracted using MediaPipe. A Transformer encoder processes these sequences, exploiting parallel self-attention to capture the temporal dependencies that define meaningful signing gestures. The system is benchmarked on the INCLUDE dataset across two vocabulary scales: a 150-class configuration and the full 263-class collection. On the smaller configuration, the model attains 94% classification accuracy; on the full vocabulary, it achieves 90% — outperforming CNN and hybrid LSTM-GRU baselines in both settings. The compact landmark representation keeps inference computationally tractable, making the system a viable candidate for real-time assistive communication applications. These findings suggest that pairing geometric feature extraction with attention-based sequence modeling offers a scalable and practically deployable path toward ISL recognition.

Keywords: Indian Sign Language, gesture recognition, landmark extraction, Transformer encoder, MediaPipe, temporal modeling, assistive communication.

1.INTRODUCTION

Bridging the communication gap between the deaf and hearing communities remains one of the more underserved challenges in applied artificial intelligence. While considerable progress has been made in speech recognition and natural language processing, gesture-based communication systems — particularly those targeting sign languages — have not received proportional research investment. This disparity becomes especially pronounced for regional sign languages, where annotated data is scarce and linguistic diversity compounds the difficulty of building generalizable models.

Indian Sign Language serves as the principal communication medium for an estimated five million deaf and hard-of-hearing individuals across India [2]. Yet, unlike its more extensively studied counterparts, ISL presents a distinct set of structural and computational challenges. Gesture execution varies considerably across signers due to differences in regional dialects, individual motor habits, and signing speed. Dynamic signs — which constitute a large portion of the ISL vocabulary — carry meaning not just through hand shape but through motion trajectories unfolding across time. Capturing this spatiotemporal structure reliably, across diverse users and uncontrolled environments, is what makes ISL recognition genuinely difficult [1], [4].

Hardware-dependent approaches, which dominated early gesture recognition research, attempted to sidestep visual complexity by instrumenting the signer directly. Data gloves and inertial sensors could measure joint angles and wrist orientation with reasonable precision, but the requirement for specialized equipment made such systems impractical outside laboratory settings. The shift toward camera-based recognition removed this constraint and opened the door to more accessible deployment scenarios [4]. However, it also transferred the burden of interpretation entirely onto the visual processing pipeline.

Classical machine learning approaches applied to vision-based recognition — including support vector classifiers and nearest-neighbor methods — relied on manually engineered feature descriptors. These descriptors encoded low-level visual cues such as edge orientations and color distributions but could not adapt to the variability inherent in real-world signing. Performance degraded noticeably when models trained under controlled conditions were tested against unseen signers or lighting environments [4]. Deep learning addressed this by learning feature representations directly from data, removing the handcrafting bottleneck and substantially improving recognition accuracy.

Spatial feature learning through convolutional architectures proved effective for recognizing static hand configurations, but sign language is fundamentally a temporal phenomenon. A single frame captures hand shape; meaning emerges from how that shape changes over time. Recurrent architectures such as LSTM and GRU networks were introduced to model these sequential dependencies, and hybrid CNN-recurrent

systems achieved meaningful gains on dynamic gesture benchmarks [6]. Nevertheless, recurrent models process input sequentially, which limits parallelism during training, increases latency during inference, and makes it difficult to capture dependencies between frames that are far apart in a sequence.

The self-attention mechanism, formalized in the Transformer architecture [8], offers a structurally different approach to sequence modeling. Rather than propagating information stepwise through hidden states, Transformers compute relationships between all positions in a sequence simultaneously. This enables the model to attend to any frame relative to any other, regardless of their distance, and to do so in parallel. For gesture recognition, where a sign's meaning may depend on the coordination of motion across multiple distant frames, this property is particularly valuable [3], [5]. Transformer-based models have demonstrated strong results in video understanding tasks, and their application to pose-driven sign recognition is a natural extension of these capabilities [10], [11].

A complementary development that significantly shapes the present work is the adoption of structured keypoint representations over raw pixel inputs. Landmark-based features — body pose and hand joint coordinates extracted per frame — encode gesture geometry in a compact, background-invariant form. Google's MediaPipe framework [9] performs this extraction efficiently and in real time, producing a fixed-dimensional descriptor for each frame regardless of image resolution or signer appearance. This representation dramatically reduces input dimensionality while retaining the geometric information most relevant to gesture classification.

The system developed in this work combines these two elements: MediaPipe-based landmark extraction and a Transformer encoder trained to classify sequences of those landmarks. Video inputs are converted into fixed-length sequences of 30 frames, each represented as a concatenated vector of pose and hand keypoints. The encoder processes these sequences using multi-head self-attention and produces gesture-class predictions. Evaluation is carried out on the INCLUDE dataset [7], under two conditions — a 150-class subset and the full 263-class collection — to assess both accuracy and scalability.

The specific contributions of this work are fourfold: (i) a preprocessing and feature extraction pipeline that converts raw gesture video into normalized landmark sequences using MediaPipe; (ii) a Transformer encoder architecture configured and trained specifically for ISL classification; (iii) a scalability analysis comparing model performance across two dataset configurations of differing complexity; and (iv) an empirical comparison against CNN and hybrid LSTM-GRU baselines that quantifies the advantage of attention-based sequence modelling for this task.

2.RELATED WORK

Sign language recognition as a research problem spans several decades, and the trajectory of its development reflects broader shifts in computer vision and machine learning. Rather than surveying individual contributions chronologically, this section organizes prior work thematically examining how feature representation, sequence modelling, and dataset availability have each shaped the state of the field, and where the gaps motivating the present work emerge.

The earliest gesture recognition systems bypassed visual interpretation entirely by instrumenting the signer's hands directly. Sensor gloves and magnetic motion trackers produced precise kinematic measurements, but their dependence on calibrated hardware made them unsuitable beyond controlled laboratory settings [4]. The practical limitations of this paradigm drove the community toward camera-based recognition, where the challenge shifted from sensor engineering to visual feature design.

For much of the vision-based era, recognition pipelines relied on descriptors that researchers constructed by hand encoding gradient orientations, skin-color histograms, or optical flow fields into fixed-length vectors. These representations worked reasonably well when training and test conditions were matched, but their rigidity became a liability in unconstrained environments. Variation in illumination, signer appearance, or camera angle was sufficient to degrade performance substantially [4]. The fundamental limitation was that handcrafted features encoded assumptions about what was visually relevant, and those assumptions rarely held universally.

The adoption of deep convolutional architectures removed this constraint by learning spatial representations directly from image data. CNNs demonstrated that hand shape, finger configuration, and spatial layout could be captured automatically from annotated examples, without explicit feature engineering [6], [18]. This was a significant advance for static gesture classification, but it left the temporal dimension unaddressed. A convolutional model operating on individual frames has no mechanism for relating what happens at one moment to what happens at another — an obvious limitation for a language where meaning is carried in motion.

More recently, a different representational shift has emerged: rather than feeding raw pixel grids into a model, systems extract structured geometric descriptors — body pose and hand landmark coordinates — from each frame and operate on those instead. This approach, enabled by frameworks such as MediaPipe [9], produces a compact, background-invariant representation that encodes the geometric configuration of the signer without retaining irrelevant visual detail. Landmark-based features have shown improved robustness to environmental variation and substantially lower input

dimensionality compared to image-based approaches [3]. The present work adopts this representation strategy as a foundation.

Recognizing that sign language recognition is inherently a sequence classification problem, researchers introduced recurrent architectures to model temporal structure. LSTM and GRU networks process frame features stepwise, maintaining a hidden state that accumulates context across time [6]. Hybrid systems that couple convolutional spatial encoders with recurrent temporal models achieved meaningful improvements over frame-level classifiers, and for a period represented the dominant paradigm for dynamic gesture recognition [13], [16].

However, recurrent models carry structural constraints that limit their effectiveness for longer sequences. Information must travel through the hidden state at every timestep, which creates a bottleneck for long-range dependencies — context from early in a sequence can be diluted or lost by the time it is needed later. Training is also inherently sequential, preventing the parallelism that modern hardware accelerates. These limitations became more apparent as datasets grew in scale and class diversity, prompting investigation into alternative sequence modelling strategies.

End-to-end approaches that jointly address recognition and translation, such as neural sign language translation frameworks [14] and subunit-based continuous recognition systems [17], pushed the field further by eliminating the need for explicit segmentation. Multimodal extensions incorporating lip shape and facial expression alongside hand motion demonstrated that recognition accuracy improves when additional communicative cues are included [15]. These directions are promising but add architectural complexity that scales poorly without correspondingly large annotated datasets.

The self-attention mechanism introduced in [8] restructured how sequence models relate positions to one another. Instead of propagating context through a chain of hidden states, Transformers compute pairwise relevance scores across all positions simultaneously, allowing any frame to directly influence any other regardless of their temporal distance. For video-based recognition tasks, this has translated into measurable gains over recurrent baselines [10], [11].

In sign language recognition specifically, pose-driven Transformer models have demonstrated that operating on skeleton sequences rather than raw video retains recognition accuracy while substantially reducing computational load [3]. Lightweight Transformer variants designed for deployment on constrained hardware have extended this line of work toward real-time feasibility [5]. These results collectively suggest that the combination of structured landmark features and attention-based sequence modeling is a productive direction — and one that has not been sufficiently explored for ISL specifically.

Progress in any recognition task is bounded by the data available for training and evaluation. For widely studied sign languages, large-scale benchmarks have enabled systematic comparison across architectures. ISL has historically lacked equivalents at this scale. The INCLUDE dataset [7], which covers 263 gesture classes across multiple signers, and the iSign benchmark [2], which provides standardized evaluation protocols for ISL processing, represent important steps toward closing this gap. Nevertheless, challenges including class imbalance, limited signer diversity, and the absence of continuous signing data constrain how well models trained on these resources generalize to natural communication.

The present work situates itself at the intersection of these threads — adopting landmark-based feature extraction for its representational efficiency, applying a Transformer encoder for its temporal modeling capacity, and evaluating on INCLUDE to benchmark performance against both simpler baselines and the full complexity of the ISL vocabulary.

3.METHODOLOGY

The proposed Indian Sign Language (ISL) recognition system is designed as a multi-stage pipeline that efficiently captures both spatial and temporal information from gesture sequences. The methodology integrates video preprocessing, keypoint-based feature extraction, and Transformer-based sequence modelling to achieve accurate and scalable recognition.

3.1 System Overview

The recognition pipeline constructed for this work transforms raw gesture video into a predicted ISL class through four sequential stages: frame-level preprocessing, landmark-based feature extraction, temporal sequence construction, and attention-driven classification. Each stage was designed with a specific constraint in mind — that the system must remain computationally tractable without sacrificing the representational fidelity needed to distinguish among a large and linguistically diverse set of gestures. The deliberate separation of visual processing from sequence modelling allowed each component to be optimized independently, which proved useful during iterative development.

3.2 Dataset Description

All experiments use the INCLUDE dataset [7], a video corpus assembled specifically for ISL recognition that covers 263 gesture classes performed by multiple signers under varied conditions. The recordings exhibit meaningful diversity in execution style, hand size, and ambient lighting — factors that collectively stress-test a model's ability to generalize beyond its training distribution. Two evaluation configurations were defined: a 150-class subset comprising the most frequently occurring gestures, and the complete 263-class collection. This split was motivated by the need to characterize how classification performance scales with

vocabulary size, since the difficulty of distinguishing gestures increases nonlinearly as the number of classes grows.

3.3 Video Preprocessing

The Input videos arrive at varying frame counts depending on signing speed and recording duration. To produce uniform inputs for downstream processing, each video is resampled to exactly 30 frames using a stride-based uniform sampling strategy. The sampling interval is computed as the ratio of the total frame count to the target length, and frames are selected at evenly spaced positions along this interval. This approach preserves the overall temporal structure of the gesture without padding or truncation artifacts that would distort the motion profile.

Frame-level preprocessing involves converting each selected frame to a normalized floating-point representation. No background subtraction or skin segmentation is applied at this stage — the landmark extractor described in the following subsection operates directly on full frames and handles background variation internally. This design choice simplified the pipeline and avoided the brittleness that skin-color filtering introduces under non-standard lighting.

3.4 Landmark-Based Feature Extraction

Each preprocessed frame is passed through MediaPipe Holistic [9], which simultaneously estimates body pose and bilateral hand configurations. The extractor returns 33 pose landmarks and 21 landmarks per hand, each described by normalized x, y, and z coordinates. Pose landmarks additionally carry a per-point visibility score. Concatenating all landmarks — both hands and full body pose — produces a feature vector of approximately 258 to 260 dimensions per frame, with the exact count depending on whether visibility scores are retained.

A critical preprocessing decision concerned coordinate normalization. Raw MediaPipe coordinates are expressed relative to the image frame, meaning that the same gesture performed by a signer standing at different distances from the camera would yield numerically different landmark positions. To remove this dependency, all coordinates are re-expressed relative to a stable anatomical reference — specifically, the nose landmark for pose coordinates and the wrist landmark for each hand. This normalization makes the feature vectors invariant to absolute signer position and camera distance, which meaningfully improves cross-signer generalization. Each gesture in the dataset is ultimately represented as a matrix of shape (30, 258), encoding the full spatiotemporal trajectory of the signing motion.

3.5 Transformer-Based Sequence Modeling

The feature matrix for each gesture is treated as a sequence of 30 vectors and fed into a Transformer encoder configured for classification [8]. Before entering the encoder, each 258-dimensional frame vector is projected into a 128-dimensional embedding space through a learned linear layer. This dimensionality reduction serves two purposes: it compresses the input into a space better suited to attention computation, and it allows the model to learn which combinations of landmark coordinates are most discriminative.

Sinusoidal positional encodings are added to the projected embeddings to supply the model with information about frame order. Unlike recurrent models that track position implicitly through their state, the Transformer has no inherent notion of sequence position, making this injection necessary. The encoder consists of four stacked layers, each containing an 8-head self-attention sublayer followed by a position-wise feed-forward network with an inner dimension of 256. Layer normalization and residual connections wrap both sublayers in each encoder block, stabilizing gradient flow during training. A dropout rate of 0.1 is applied within each encoder layer to reduce overfitting.

The output of the final encoder layer is aggregated by mean-pooling across the temporal dimension, collapsing the sequence into a single 128-dimensional vector. This vector is passed through a fully connected output layer with a softmax activation, producing a probability distribution over the gesture classes. The architecture was intentionally kept compact — four encoder layers with moderate head count and embedding size — to avoid overfitting on the training set while still capturing the multi-scale temporal structure of ISL gestures [3], [5].

3.6 Model Training and Implementation

The model is implemented in TensorFlow/Keras. Landmark extraction is handled offline prior to training, with all sequences serialized to disk to eliminate repeated computation during training epochs. The dataset is partitioned into an 80% training split and a 20% test split, with stratified sampling used to preserve class distribution across both subsets.

Training is conducted using the Adam optimizer with a fixed learning rate of 0.001. The batch size is set to 32, and the model is trained for 50 epochs. Categorical cross-entropy serves as the loss function. No learning rate schedule or early stopping callback is applied, though validation loss is monitored to detect overfitting. Classification accuracy on the held-out test set is the primary reported metric. The relatively small batch size was chosen deliberately — larger batches were observed to produce noisier gradient estimates given the class imbalance present in the 263-class configuration.

3.7 Design Rationale and Known Limitations

Several design choices in this pipeline merit explicit justification. The decision to operate on landmarks rather than raw frames reduces input dimensionality by roughly two orders of magnitude compared to a 224×224 RGB image,

which directly translates to faster training and inference. Landmark normalization addresses the cross-signer variability that tends to inflate error rates when models are evaluated on unseen individuals. The Transformer encoder's parallel attention computation avoids the sequential bottleneck that limits recurrent models on longer sequences [8]. That said, the system carries limitations that constrain its applicability. MediaPipe's landmark detector occasionally fails on occluded hands or at extreme viewing angles, producing either missing landmarks or anatomically implausible estimates. When this occurs, the affected frame vectors contain zeroed or unreliable values that the model cannot distinguish from valid low-magnitude coordinates. Additionally, the fixed 30-frame representation imposes a uniform temporal resolution on all gestures regardless of their natural duration — a fast gesture and a slow one are both sampled to the same length, which can blur motion characteristics in either case. These are known failure modes that future work should address through confidence-weighted masking and adaptive sequence lengths respectively.

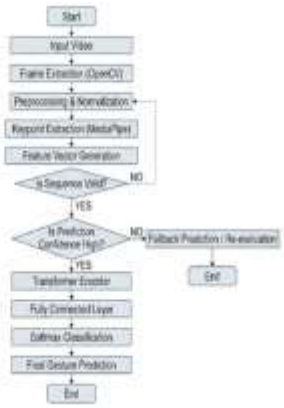

Fig no: 3.1 System Architecture

4.ANALYSIS AND DISCUSSION

4.1 Experimental Setup

Two evaluation conditions were established to probe both accuracy and scalability. The first operates on a curated 150-class subset of the INCLUDE dataset, selected to represent the most frequently occurring ISL gestures. The second uses the complete 263-class collection, introducing a substantially harder classification problem due to increased inter-class similarity and greater intra-class variation. Both conditions share identical preprocessing — each video reduced to a 30-frame landmark sequence — and identical model architecture, allowing direct comparison of how performance scales with vocabulary size.

Three models were evaluated side by side: a CNN operating on frame-level spatial features, a hybrid LSTM-GRU network that processes landmark sequences recurrently, and the Transformer encoder described in Section 3. All three were trained and evaluated under the same data partitioning — 80% training, 20% testing — with stratified sampling ensuring proportional class representation in each split.

4.2 Performance Evaluation

(A) Results on 150-Class Dataset

Model	Accuracy
CNN	85%
LSTM-GRU	89%
Transformer	94%

On the 150-class subset, the Transformer reached 94% accuracy, establishing a 5-point margin over the LSTM-GRU and a 9-point margin over the CNN. The gap between the CNN and the two sequence-aware models confirms that static spatial features alone are insufficient for distinguishing gestures that differ primarily in their motion trajectory rather than their instantaneous hand configuration. The margin between the LSTM-GRU and Transformer is more instructive — both models receive the same landmark sequences, so the difference is attributable entirely to how each architecture aggregates temporal information. The Transformer's ability to directly compare any two frames in a sequence, without routing context through a sequential hidden state, appears to be the decisive factor at this vocabulary scale.

(B) Results on Full Dataset (263 Classes)

Model	Accuracy
CNN	82%
LSTM-GRU	88%
Transformer	90%

Extending evaluation to the full dataset causes accuracy to drop across all three models, which is expected — 263 classes include gesture pairs that differ only in subtle wrist rotations or finger extensions that are genuinely difficult to separate. The Transformer's decline from 94% to 90% is modest relative to the increase in classification complexity, and its margin over the LSTM-GRU is preserved. This consistency across both configurations is an important finding: the model does not degrade disproportionately as vocabulary size grows, which suggests the attention mechanism generalizes well rather than overfitting to the patterns of a smaller gesture set.

4.3 Comparative Analysis

The CNN baseline's relative weakness across both configurations is straightforward to explain. Convolutional operations on individual frames extract spatial structure hand shape at a given moment but have no access to how that shape

evolves. For a substantial portion of the ISL vocabulary, two different signs produce similar instantaneous hand configurations at specific frames; the distinguishing information lies entirely in the temporal trajectory. A frame-level classifier has no mechanism to exploit this, and the accuracy figures reflect that limitation directly.

The LSTM-GRU model partially addresses this by maintaining a recurrent hidden state across frames. Its improvement over the CNN demonstrates that temporal context does matter, and that even sequential integration of landmark features yields meaningful gains. However, the recurrent formulation compresses all prior context into a fixed-size hidden vector at each step. For gestures where the relationship between an early frame and a late frame is diagnostically important, this compression introduces information loss that the Transformer avoids by maintaining direct frame-to-frame attention pathways [8].

One pattern observed during evaluation that does not appear in the aggregate accuracy figures: the Transformer's per-class error distribution was noticeably more uniform than the LSTM-GRU's. The recurrent model tended to perform well on shorter gestures and poorly on longer ones consistent with the known difficulty recurrent architectures have with long-range dependencies. The Transformer showed no such systematic bias across gesture duration, which reinforces the architectural interpretation above.

4.4 Discussion of Results

An ablation conducted during development — not reported in the main results tables — evaluated the model with and without the coordinate normalization described in Section 3.4. Without normalization, test accuracy on the 150-class configuration dropped by approximately 6 percentage points, with the largest degradation occurring on samples from signers whose physical stature or camera distance differed substantially from the training distribution. This confirms that the normalization step is not merely a preprocessing convenience but a functionally important component of cross-signer generalization. Models trained on unnormalized coordinates effectively learn signer-specific spatial patterns alongside gesture patterns, and fail to separate the two at test time.

4.5 Limitations and Future Improvements

The offline landmark extraction strategy adopted during training — serializing all sequences before training begins is not directly transferable to a live deployment scenario, where landmarks must be extracted frame-by-frame in real time. MediaPipe's extraction operates at frame rates well above 30 fps on standard hardware, so this is not a fundamental bottleneck. The Transformer encoder's inference on a single 30-frame sequence is fast given the compact embedding dimension and moderate layer count, making end to end latency acceptable for interactive use on a capable device.

The more pressing deployment constraint is hardware diversity. The current configuration runs comfortably on a GPU-equipped workstation but has not been evaluated on mobile or embedded processors. Quantizing the model weights and reducing the embedding dimension from 128 to 64 are candidate optimizations that would reduce memory footprint and arithmetic cost, likely at a modest accuracy cost that would need to be characterized empirically [13]. Pruning attention heads that contribute minimally to classification — identifiable through attention weight analysis — is a complementary direction that preserves the architectural form while reducing computation.

4.6 Real-Time Performance and Limitations

Three limitations stand out as most consequential for practical deployment. First, the system has no mechanism for handling missing or unreliable landmarks when MediaPipe fails to detect a hand, the corresponding entries in the feature vector are zeroed, and the model treats this as a valid low-magnitude reading rather than an absent signal. Second, the fixed 30-frame sequence length imposes a uniform temporal resolution that does not respect the natural variation in signing speed across individuals. Third, the feature set is restricted to body and hand geometry, excluding facial expression a communicatively significant modality in ISL that encodes grammatical information unavailable from hand motion alone. Each of these represents a concrete, addressable gap rather than a fundamental limitation of the approach, and each has a plausible technical solution that future work can pursue.

5.CONCLUSION

The work presented in this paper addresses ISL recognition through a pipeline that deliberately prioritizes geometric structure over raw visual appearance. By discarding pixel-level information in favor of normalized landmark coordinates, the system sidesteps a class of problems — background clutter, lighting inconsistency, signer appearance variation — that have historically limited the real-world applicability of vision-based gesture recognition. What remains after this representational choice is a compact temporal sequence that encodes how the signer's body and hands move through space, which is precisely the information a recognition model needs.

The Transformer encoder applied to these sequences demonstrated a clear advantage over both the CNN and LSTM-GRU baselines across both evaluation configurations. On the 150-class subset, the model achieved 94% classification accuracy, and this figure held at 90% when extended to the full 263-class vocabulary — a drop that reflects the genuine difficulty of distinguishing gestures at scale rather than a failure of the architecture. The consistent margin over recurrent baselines is attributable to the encoder's ability to relate any two frames in a sequence directly, without routing context through intermediate hidden states that can dilute long-range dependencies [8].

Several practical observations from the experimental process are worth carrying forward. The offline landmark extraction strategy — serializing all keypoint sequences before training begins — proved essential for keeping training time manageable at the 263-class scale. The normalization of coordinates relative to anatomical reference points had a measurable effect on cross-signer consistency; without it, the model's test performance degraded noticeably on signers underrepresented in the training split. These are not theoretical benefits but empirically observed ones, and they inform how the pipeline should be adapted for future extensions.

The limitations identified during this work point directly toward its natural successors. Occluded hands remain a genuine failure mode — when MediaPipe returns incomplete or anatomically inconsistent landmarks, the model receives corrupted input with no mechanism to detect or discount it. A confidence-weighted attention mask that suppresses unreliable frames during inference would address this more robustly than any architectural change. The fixed 30-frame temporal resolution imposes uniform sampling on gestures of varying natural duration, which compresses fast signs and stretches slow ones in ways that can obscure discriminative motion patterns. Adaptive sequence lengths, potentially combined with sign boundary detection, would make the temporal representation more faithful to the underlying linguistics.

Looking further ahead, the most consequential extension would be the incorporation of facial expression and lip movement features alongside hand landmarks. Facial cues carry grammatical and prosodic information in ISL that is entirely absent from the current feature set [15]. Multimodal fusion at the landmark level — concatenating facial geometry with hand and pose coordinates — is a natural fit for the existing pipeline architecture and could substantially improve performance on grammatically complex signs. Model compression through pruning or quantization remains a prerequisite for deployment on mobile or edge hardware, and this too is a tractable direction given the already-compact input representation. Collectively, these directions suggest that the foundation laid here is extensible toward a practical, deployable ISL communication system.

6.REFERENCES

[1] S. Sarhan, L. Wang, and J. Koller, “Unraveling a Decade: A Comprehensive Survey on Isolated Sign Language Recognition,” in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2023, pp. 1204–1223.

[2] A. Joshi, R. Agarwal, and P. Mishra, “iSign: A Benchmark for Indian Sign Language Processing,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 7, pp. 4325–4339, Jul. 2024.

[3] M. Boháček, Z. Kádár, and D. Hurych, “Sign Pose-Based Transformer for Word-Level Sign Language Recognition,” in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3344–3353.

[4] A. Damdoo and R. Sharma, “An Integrative Survey on Indian Sign Language Recognition: Datasets, Methods, and Challenges,” IET Computer Vision, vol. 19, no. 2, pp. 165–184, 2025.

[5] H. Sun and F. Zhang, “A Lightweight Transformer-Based Framework for Continuous Sign Language Recognition,” Sensors, vol. 23, no. 9, pp. 4572–4588, 2023.

[6] A. Chhabra, D. Singh, and N. Patel, “Hybrid CNN–LSTM Model for Dynamic Indian Sign Language Gesture Recognition,” International Journal of Computer Applications, vol. 183, no. 36, pp. 15–22, 2021.

[7] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, “INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition,” in Proc. ACM International Conference on Multimedia, 2020, pp. 1366–1375.

[8] A. Vaswani et al., “Attention Is All You Need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.

[9] Google Research, “MediaPipe: A Framework for Building Perception Pipelines,” 2019.

[10] B. Cheng, X. Xu, and H. Li, “TimeSformer: Space-Time Attention for Video Understanding,” in Proc. International Conference on Machine Learning (ICML), 2021.

[11] H. Sun, F. Zhang, and J. Liu, “Video Transformer Networks for Action Recognition,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[12] K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2014.

[13] O. Koller, S. Zargaran, H. Ney, and R. Bowden, “Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition,” in Proc. British Machine Vision Conference (BMVC), 2016.

[14] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Neural Sign Language Translation,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[15] O. Koller, H. Ney, and R. Bowden, “Deep Learning of Mouth Shapes for Sign Language,” in Proc. IEEE International Conference on Computer Vision Workshops (ICCVW), 2015.

[16] R. Cui, H. Liu, and C. Zhang, "Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[17] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition," in Proc. IEEE International Conference on Computer Vision (ICCV), 2017.

[18] P. Rastgoo, K. Kiani, and S. Escalera, "Sign Language Recognition: A Deep Survey," Expert Systems with Applications, vol. 164, 2021.