International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-14, Issue-1, January 2026
DOI: https://doi.org/10.55524/ijircst.2026.14.1.5
Article ID IJIRE-1416, Pages 34-42
www.ijircst.org

Sign Language Recognition Using Deep Learning: A Systematic Review of Models and Approaches

Sruthi SJ[1], Evaniya Anna Suvi[2], Rejeena J Jajin[3], Annfino Jagan[4], and Shakhy PS[5]

[1, 2, 3, 4] B.Tech Scholar, Department of Computer Science and Engineering, Marian Engineering College, Trivandrum, India
[5]Assistant Professor, Department of Computer Science and Engineering, Marian Engineering College, Trivandrum, India

Correspondence should be addressed to Sruthi S J; sruthisj6162@gmail.com

Received: 23 November 2025 Revised: 10 December 2025 Accepted: 24 December 2025

ABSTRACT- SLR has evolved as one of the most important areas in human–computer interaction and assistive communication technologies. With the rapid development of deep learning, SLR systems have evolved from traditional handcrafted features to highly efficient data-driven models capable of grasping complex spatial and temporal patterns in sign sequences. A wide variety of recent works have investigated different approaches that range from CNNs and recurrent architectures to GCNs for skeletal modelling and state-of-the-art transformer frameworks for long-range sequence understanding. Besides, multimodal systems that incorporate RGB, depth, skeletal information, and radio-frequency signals demonstrate enhanced robustness under difficult real-world conditions. This survey provides an in-depth review of modern advances in SLR by underlining methodological novelties, commonly used datasets, architectural enhancements, and corresponding performance results. Shared challenges regarding signer variability, limited diversity in datasets, occlusions, and constraints on real-time processing are discussed in detail. The survey concludes by underlining emerging trends and future research directions oriented to the development of scalable, accurate, and context-aware SLR systems that can be effectively used in practical assistive applications.

KEYWORDS- Sign Language Recognition, Deep Learning, CNN, GCN, Transformers, Pose Estimation, Multimodal Fusion, Continuous SLR, Word-Level SLR, Human–Computer Interaction.

I. INTRODUCTION

Sign languages are developed, natural languages that are entirely visual. They are extensively utilized within Deaf communities internationally. Deaf sign languages employ sensory channels simultaneously. These encompass hand configuration, motion, direction, placement and facial gestures. The primary cause that conventional automatic speech recognition systems cannot be modified for sign languages is the absence of signals, in sign languages. The goal of the Automatic Sign Language Recognition System is to decode these cues and convert them into spoken or written words [1] [2] [3] [4] [5] [6].

Advancements in computer vision and deep learning have swiftly transformed the landscape of SLR research. Convolutional Neural Networks excel at extracting features from images effectively identifying hand configurations and movements facial expressions and subtle articulations that are challenging to model with manually designed features [7] [8]. Recurrent Neural Networks, particularly LSTM and GRU variants have proven effective in modeling the progression of sign production and detecting minute temporal variations essential, for recognizing dynamic gestures and sign sequences. [9], [10].Additionally techniques based on graph and pose representations have become a focus of investigation. Graph Convolutional Networks (GCNs) which operate directly on graph structures derived from joints efficiently manage the temporal dynamics of gestures along with background interference and variation [11] [12]. Transformer networks have advanced SLR and sign language translation by utilizing self-attention to accurately model long-range dependencies, in inter-channel gestures and coarticulation [13] [14], [15].

In spite of these improvements several obstacles continue to impede the effectiveness of SLR systems. These obstacles encompass variability and domain transfer/shift involving aspects like physiology, sign style, speed of movement, perspective, attire and recording conditions among elements. [16] Ongoing SLR learning poses a greater difficulty because natural sign languages lack distinct boundaries due, to coarticulation movement epenthesis and transitional gestures. [17] [18] [19].

An additional difficulty in this area is the availability of annotated data particularly, for less resourced sign languages. Creating a quality sign language dataset demands knowledge for part annotations and multi-channel synchronization, which results in high costs and slow data collection. Consequently, certain sign languages lack datasets for the efficient training of deep learning models [20][21][22][23][24]. Additionally, elements like blocked hands, rapid movement, background distractions and poor lighting complicate recognition efforts because of environments [25]. Lastly challenges persist for SLR in video scenarios due, to the processing power needed for advanced neural models. Transformer networks, high-resolution video encoders, and multi- stream fusion architectures are quite resource-intensive and therefore can't be deployed on mobile and embedded platforms for assistive usage scenarios [26], [27], [28]. Therefore, efficient and low-latency architectures for SLR are being researched.

II. LITERATURE SURVEY

The research on SLR has grown quickly in the last ten years because of the recent advancements in deep learning and increased demand for inclusive communication systems in society. Most earlier attempts at SLR were based on handcrafted features that were highly susceptible to environmental changes and signer variations. With the arrival of deep learning, much more robust approaches emerged that are capable of automatically extracting spatial–temporal patterns, improving gesture segmentation, and increasing the overall recognition accuracy. Multiple architectural paradigms have since been investigated by researchers, including convolutional neural networks, recurrent sequence models, graph-based skeletal learning, and transformer frameworks.

This literature review consolidates major contributions across these domains, putting in perspective their methodological novelties, benchmark performance, datasets, and application constraints. The survey will present an overview of how SLR models have evolved, the strengths and limitations of existing techniques, and also the trends that will guide future research by reviewing key works covering isolated, word-level, and continuous signing. The aim of this section is to provide a structured understanding of how SLR solutions have progressively evolved toward more accurate, generalizable, and context-aware recognition systems capable of operating within real assistive environments.

A. CNN-Based Approaches

Early works on SLR with deep learning mainly relied on 'Appearance-based Methods' and were based on 2D and 3D Convolutional Neural Networks (CNNs), incorporating spatial and short-term spatiotemporal information for feature extraction from RGB videos [7], [8], [29], [30], [31], [32], [33], [34]. These methods allow automatic learning of hierarchical visual representation based on handshape configuration, local movement, and upper body pose without requiring any hand-crafted features. 2D CNNs concentrate on spatial pattern recognition at the frame level, and 3D CNNs improve upon it with simultaneous consideration of motion information on a frame-by-frame basis. To improve discriminative abilities, multi-stream CNN architectures were formulated with dedicated branches for regions of interest like hands, face, and upper body, eventually combining these streams' outputs with a common prediction task [8], [30]. The structured knowledge of sign languages allows simultaneous focus on detailed articulation based on hands and additional non-manual information based on facial expressions and upper body configuration.

Appearance-based CNNs show excellent performance on isolated sign languages recognition, as they receive an input clip with a single sign and well-defined temporal bounds and with limited co-articulation. CNNs' capabilities on local spatial and short-term temporal information might be adequate for classification in these conditions. Nevertheless, these methods have limitations when confronted with more realistic challenges. Variational factors on illumination, viewpoint, and signer's personal appearances might affect the extraction of spatial features. Moreover, with no explicit modeling outside short-term temporal windows, these methods' capabilities might be limited on continuous sign languages. Consequently, these challenges made subsequent investigations incorporate pose-based models, recurrent models, and transformers.

Figure 1: A typical CNN-based framework for sign language recognition is depicted below, showcasing the parallel CNN branches that the model uses to handle the visual/sensory inputs, along with the merging of the features and classification layers, representing the spatial as well as the short spatiotemporal properties that are captured using CNNs for isolated sign recognition.

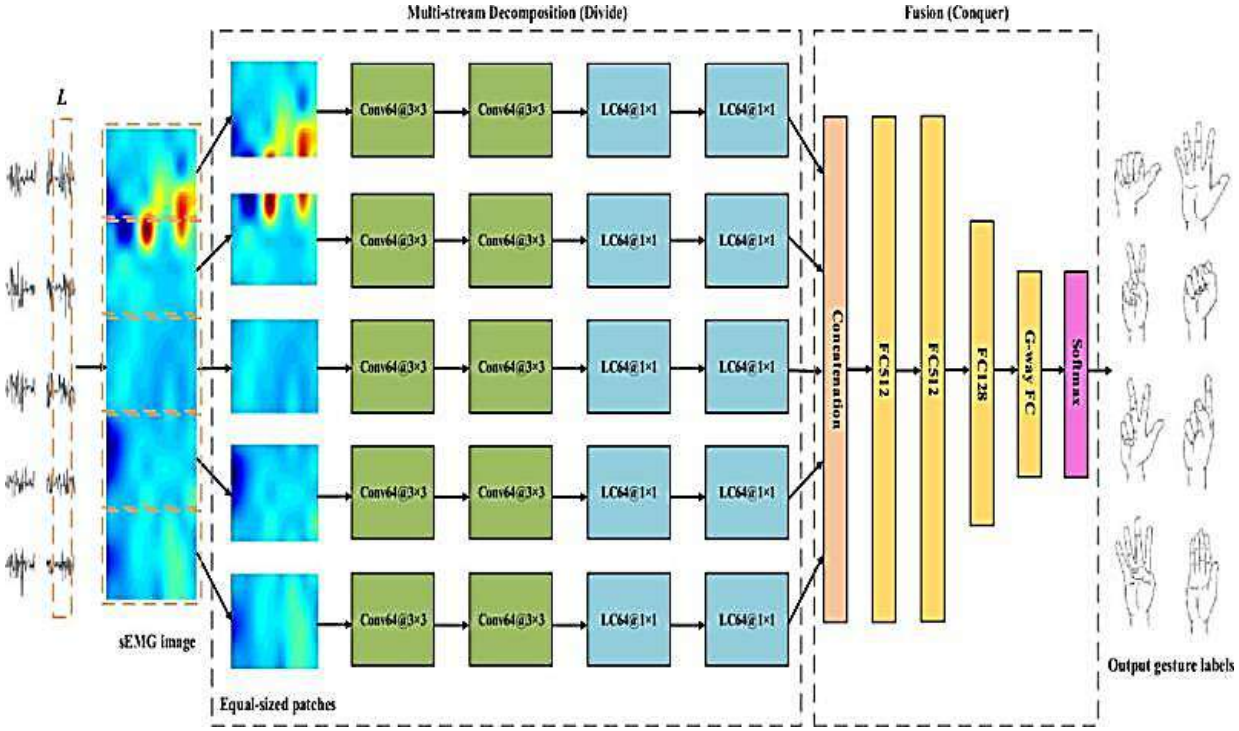

Figure 1: CNN-based multi-stream sEMG feature extraction and fusion architecture for gesture recognition.

B. RNN/BiLSTM-Based Methods

In SLR, especially, the role of sequence modeling is of great importance, as it allows the learning of long-term temporal dependencies beyond the short-range patterns captured by the CNN. This is done in a similar way to the strategy of combining CNN-based feature extractors with recurrent architectures, such as LSTMs and GRUs, to enable the tracking of motion trajectories and evolving articulations across extended sign sequences [9], [17], [35], [36], [37], [38], [39]. Hybrid CNN-RNN models are thus particularly effective under continuous recognition scenarios, in which signs unfold over time with variable duration and transitional movements. The maintenance of memory states that encode past visual information helps recurrent networks tell apart signs with similar visual appearances but distinct temporal signatures, which improves recognition accuracy and robustness against intra-signer variability.

In contrast to recurrent architectures, TCNs have become increasingly popular in modeling sequential dynamics through hierarchical stacks of dilated convolutions. As a simple, yet powerful architecture, TCNs exhibit advantages in parallelized computation, stable gradients, and flexible receptive fields that can grow exponentially with network depth. By being incorporated into CTC or encoder-decoder frameworks, TCN-based models are able to learn alignment between input video frames and corresponding gloss sequences with no explicit boundary annotations. A competing alternative to RNN-based sequence models in state-of-the-art SLR, TCNs can jointly model temporal structure and alignment in an end-to-end and computationally-efficient manner.

C. GCN-Based Skeletal Approach

The recent developments in pose estimation tools like OpenPose and MediaPipe have led researchers working on SLR focus on pose-based methods. The method uses structured coordinates representing joint locations instead of working with raw pixel intensities. By identifying 2D/3D key points representing hands, arms, and upper body regions, SLR methods based on pose estimation make recognition less dependent on factors like clothing and background complexity. It becomes feasible for the recognition system to concentrate on sign dynamics and articulation because working with skeletons significantly reduces dimensions compared to processing the entire frame size of an input image. As a result, pose estimation methods can be very efficient.

Figure 2 shows a pose-based transformer architecture for the prediction of sign language glosses. In this image, the use of skeletal key points derived from video frames, represented by attention mechanisms that focus on spatial-temporal relationships, emphasizes the insensitivity of skeletal learning to backgrounds and light conditions.

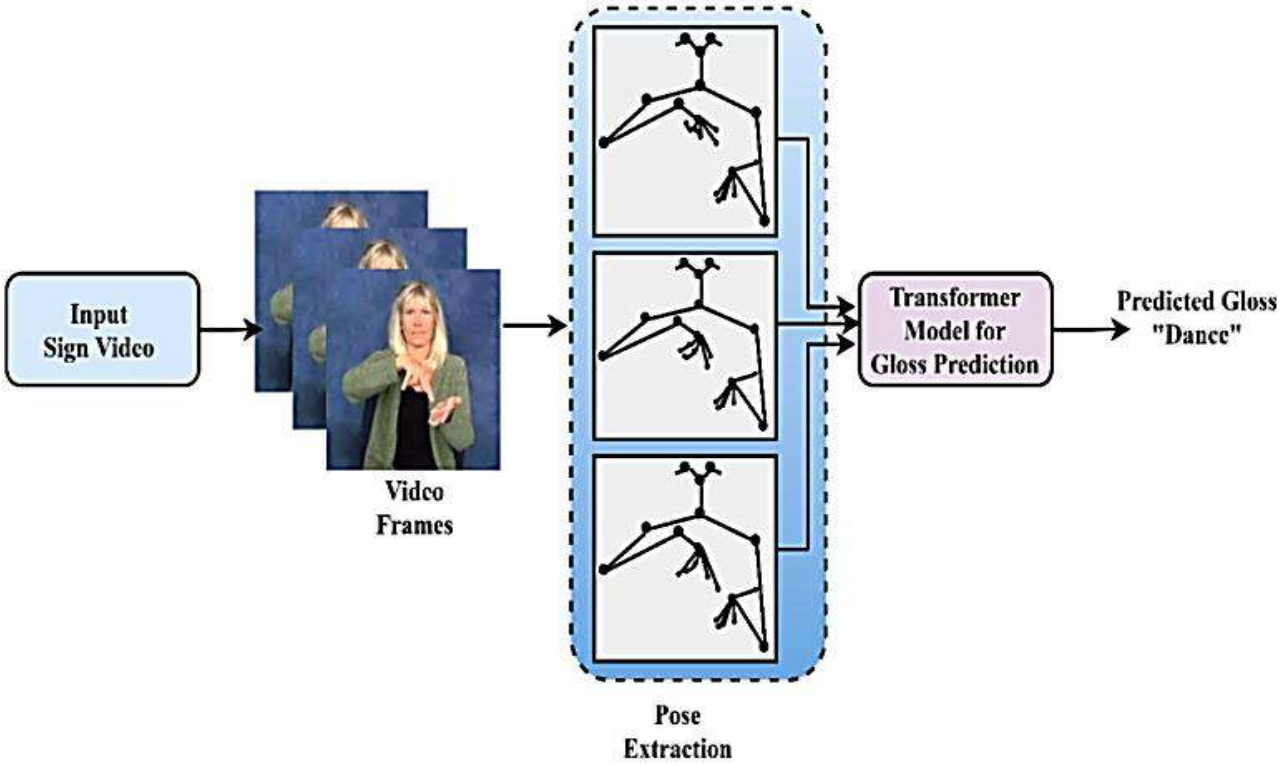

Figure 2: Pose-based transformer architecture for sign language gloss prediction from video.

Graph Convolutional Networks (GCNs), nowadays, have been recognized as the preeminent representation technique for modeling skeletal data within SLR. By representing joints as graph vertices with anatomical graph edges connecting them, GCNs seamlessly encode spatial and temporal information within consecutive frames as a result of anatomical knowledge [10], [40], [41], [42], [43], [44]. GCNs learn attributions of joints representing difference among poses with comparable global movement but divergent hand trajectory and fingeraghan configuration. Moreover, GCNs seamlessly address challenges associated with noise and missing joints, who commonly constitute limitations within pose estimation methods, making them apt tools for uncontrolled settings.

Because of their resilience against illumination changes, background presence, and motion caused by cameras, pose-based GCN methods have been recognized as an efficient alternative solution compared with conventional appearances for SLR.

D. Transformer-Based Models

Transformers have increasingly emerged as a cornerstone for state-of-the-art SLR research because they have shown efficacy at modeling global and distant dependencies without incorporating recurrent computations. Based on these benefits obtained with self-attention mechanisms, more global modeling and understanding of complex sign language sequences have been achieved without limitations on distant frames [11], [12], [45], [46], [47], [48]. Variations based on Vision Transformers, video transformers, and CNN-Transformer architectures have shown improvements with global understanding abilities via inclusion of spatial knowledge and motion-aware attributes. These architectures have maintained state-of-the-art performance on various SLR benchmark datasets due to their merits pertaining to malleability, extendibility, and efficient combination of multi-channel sources like RGB, Depth, Optical Flow, and Skeletal.

The main advantage that attention-based models have can be attributed to their capability to focus selectively on the most informative aspects of a signing sequence. It becomes possible with self-attention that allows it to focus selectively on important frames, paths, and articulations, as well as grammatical and semantic aspects, on one hand. At the same time, it captures equally well non-manual signals, like facial expressions and head movements, which assume equal importance on the grammatical and semantic understanding of sign languages. The coarticulation, ambiguous boundaries, and overlap signals are remarkably well processed with transformer architectures, and thus they form a strong foundation for next-generation SLR systems.

Figure 3 describes the internal functioning of the Transformer architecture. The overall architecture description of the encoder and decoder using multi-head attention along with the feed-forward layer is presented in Figure 3 (A), whereas Figure 3 (B) represents the scaleddot product attention layer that captures the interaction among the query, key, and value vectors.

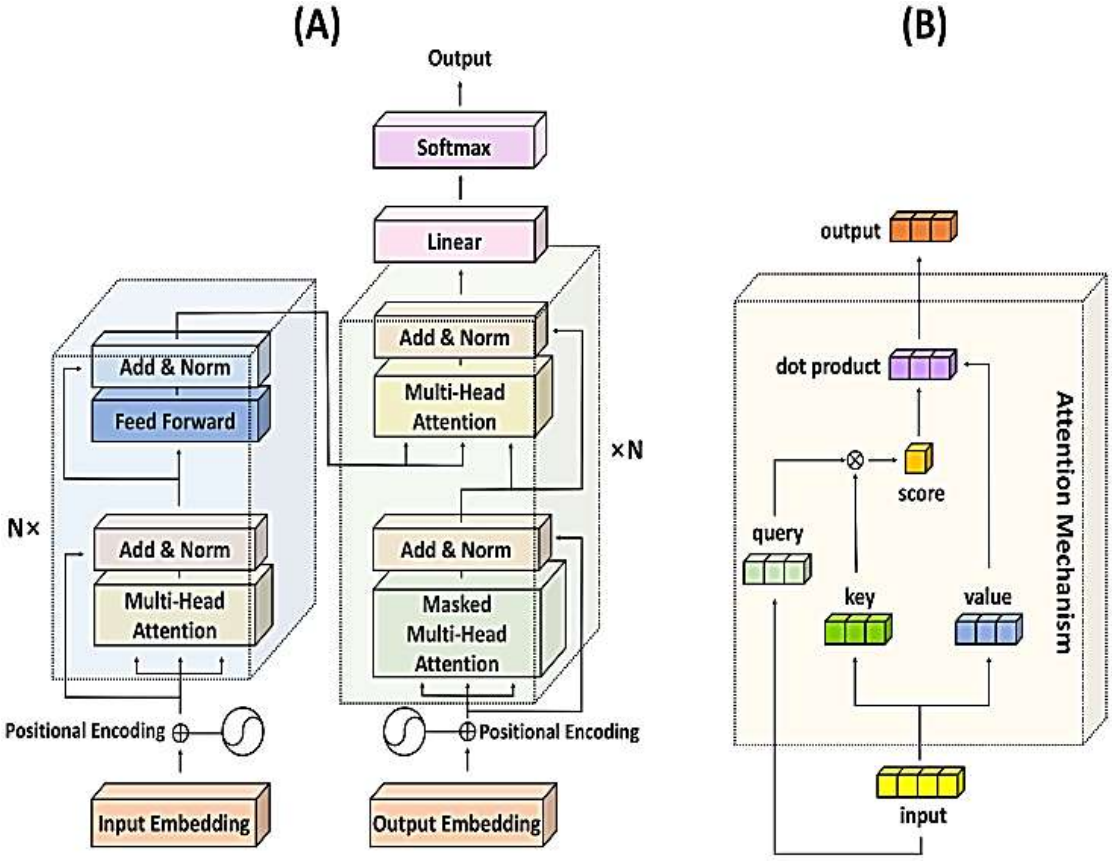

Figure 3: (A) Encoder–decoder structure of the Transformer with multi-head attention, feed-forward layers, and positional encoding. (B) Scaled dot-product attention mechanism illustrating the interaction between query, key, and value vectors.

E. Approaches Based on Fusion and Multimodality

Traditional Single-Modality SLR methods have been remedied with an innovative perspective brought about through Multimodal SLR methods. These models have incorporated multiple sources of complementary information, ranging from traditional sources like RGB images, Depth images, Optical flow images, and Key points images, and have gone a step ahead and incorporated Information from IMU. All these sources provide information about different aspects of sign interpretation. While RGB images provide information about the rich attributes of signing, Depth images form 3D structure and resolve ambiguities associated with overlapping limbs. Optical flow images depict motion attributes, Key points convey joint articulation information, and IMU signals are associated with

orientation and acceleration.

Efficient fusion techniques are critical for effectively utilizing multimodal inputs. Some research papers have explored fusion at the input level, feature level, and decision level. The input level fusion technique stacks raw modality features for simultaneous modeling; feature level fusion involves aggregating feature extraction from separate branches at the modality level. Decision level fusion integrates predictions from multiple classifiers. Through various architectures such as CNNs, RNNs, GCNs, and transformers, it has been made amply clear that fusion can significantly improve recognition accuracy and make an SLR system more resilient and apt for real-world deployments [13], [49], [50].

Figure 4 represents the structural design of a pose-based representation system for signs. Figure 4 (a) indicates spatial junction grouping, Figure 4 (b) represents hand-centered refinement layers, and Figure 4 (c) represents spatiotemporal skeleton graph building, again emphasizing multimodal fusion benefits in increasing robustness of sign detection performance.

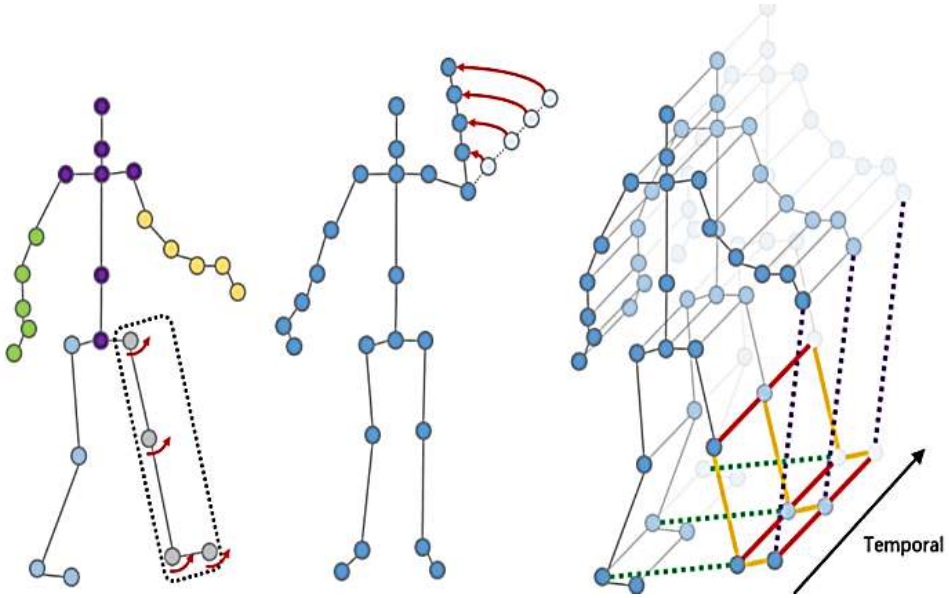

Figure 4:(a) Spatial joint grouping, (b) hierarchical hand-focused refinement, and (c) spatiotemporal skeleton graph construction used for pose-based sign representation.

F. Datasets Used in Literature

Large, annotated datasets have been a driving force behind these advances in SLR, offering the scale and variety that modern deep-learning models require. Core resources include RWTH-PHOENIX-Weather and its extension PHOENIX-2014T, central benchmarks for continuous German Sign Language and drivers of recent advances in both sequence modeling and translation tasks [2]. For American Sign Language, MS-ASL introduced a large-scale dataset with thousands of classes, facilitating research into high-vocabulary recognition and generalized representation learning [3]. Similarly, WLASL contributed a diverse, multi-signer dataset in the area of isolated sign recognition, displaying rich signer variability in terms of appearance, speed, and articulation patterns [4]. In addition to these larger corpora, an emerging set of regional datasets has expanded linguistic coverage for Chinese, Indian, Turkish, Spanish, and Arabic sign languages, reducing dependence on resource-rich languages and thus helping to support fairness and cross-linguistic generalization [5], [20], [21], [22], [23], [24].

These datasets indicate a shift from early, small-scale laboratory collections to realistic, in-the-wild benchmarks. Indeed, modern SLR corpora increasingly include natural variations of lighting, cluttered or dynamic backgrounds, signer diversity, and unconstrained camera arrangements. It is such ingredients that provide real-world signing conditions, promoting models robust beyond the carefully created, laboratory-controlled environment. It follows that very large, linguistically diverse datasets have become crucial for benchmarking of SLR architectures and driving advances in the direction of continuous recognition, translation, and multimodal fusion.

Major datasets referred to throughout SLR studies include in the below table 1:

Table 1: Comparison of benchmark sign language datasets based upon language, task type, size, and input modality

Dataset	Language	Type	Size	Modality
WLASL	ASL	Word-level	21k+ videos	RGB
MS-ASL	ASL	Word-level	25k videos	RGB
Phoenix 2014T	German SL	Continuous	9h	RGB
ASLLVD	ASL	Isolated	3.3k signs	RGB
INCLUDE	Indian SL	Isolated	Medium	RGB
HowToSign	ASL	Continuous	Large	RGB + Pose

G. Gaps in Literature Summary

SLT can be viewed as an extension with considerably greater ambitions than traditional SLR because it aims at translating entire sign videos directly into spoken sentences. Contemporary SLT models usually employ encoder-decoder frameworks with attention modules and

transformers designed to focus on capturing intricate spatiotemporal relationships among sign sequences of arbitrary lengths [18],[19],[46]. While the encoder might be composed of CNNs, RNNs, GCNs, or even transformers that focus on encoding a high-level representation of sign videos based on the visual inputs, the role of the decoder would be producing grammatically correct spoken sentences.

Despite these advances, SLT still proves to be a very hard task because of some intrinsic properties that exist due to the difference in structure between sign languages and spoken languages. The sign languages incorporate very intricate spatial grammars, classifiers, role-shifting, and simultaneous manually and non-manually tracked gestures, making it very hard to map these into the linear word structure of spoken languages. Variations in facial expressions, head movement, and signing space are meaningful and should be carefully incorporated into the translation task by the learning model. Despite these difficulties, advances made in alignment learning, attentions, and large multilingual corpora have made it easier and more efficient. Recent works based on transformers have made very encouraging breakthroughs, indicating that advances made in learning multimodal representations will narrow the gap between visual-communicative and natural languages.

III. PROPOSED METHODOLOGY

This survey follows a structured and transparent methodological framework to make the review of existing SLR research systematic, comprehensive, and unbiased. The goal is to provide a comprehensive overview of modern deep-learning methods, performance, and future research trajectories. To accomplish this, the methodology will follow several key stages, including the identification of literature selection criteria, categorization strategies, comparative evaluation, and analytical synthesis.

A. Research Framework

The proposed methodology follows a structured research framework that is better suited to clearly and systematically present the evolution in the SLR domain. The survey covers only those studies addressing deep-learning-based architectures, including CNN, RNN, GCN, transformer, and more recently, multimodal systems. Studies have been collected from valid scientific sources and subsequently filtered for their contribution to automatic sign interpretation, experimentations based on approved datasets, and architectural novelty. By narrowing the scope to methods developed around deep learning and widely referenced datasets of sign languages, the framework ensures that only impactful contributions relevant to the technical aspects are evaluated. The final set of research works reflects the chronological evolution of SLR methods from early vision-based models to transformer-driven and multimodal fusion frameworks.

B. Criteria of Categorization

After gathering the relevant literature, the studies were categorized on consistent parameters to facilitate comparative understanding. Each work was first classified according to the learning model employed, thus allowing clear separation between CNN, GCN, Transformer, and hybrid architectures. Further categorization was done based on the type of modality used in the input stream, including RGB frames, pose-based skeletal data, depth images, optical flow, and RF signal-based sensing. The task domain further helped in categorizing the methods developed for either isolated gestures, word-level recognition, or continuous signing. The approach to temporal modeling was also noted, whether achieved through recurrent units, temporal convolutions, attention mechanisms, or cross-modal fusion. These are the categorization criteria that ensure works of similar intent, computational design, and learning philosophy are analyzed against each other for a more meaningful and uniform comparison.

C. Performance Appraisal

Performance appraisal constitutes the core analytical stage of the proposed methodology. Each selected work is assessed against the reported results about benchmark datasets, model training strategies, and accuracy-related indicators. This assessment largely highlights general recognition accuracy, Top-k correctness, and quality in temporal alignment, while also considering Word Error Rate in continuous sign prediction or sequence-to-text translation metrics such as BLEU or ROUGE scores. Practical aspects, like model inference speed, stability under complex backgrounds, or robustness related to lighting variations and signer-specific differences, are also explored to gain insight into the readiness of these systems for real-world deployment. Due to this standard uniformity in evaluation, the appraisal lucidly provides insight into how various architectures perform under changing data conditions, input modalities, and application constraints.

D. Discussion and Key Findings

The methodology closes with a structured discussion synthesizing research outcomes, underlining notable trends in performance, and pointing out recurrent limitations. The comparison study shows that transformer-based systems are the most promising on large-scale datasets due to their strong temporal encoding capabilities, while GCN-based skeletal models maintain stability under conditions of low visibility and cluttered scenes. Multimodal fusion shows consistently better robustness, reducing signer-dependency by including pose, depth, and RF-sensor streams in addition to RGB imagery. Observed patterns emphasize maturity and diversification but also point to challenges concerning dataset scarcity, signer variability, continuous signing segmentation, and real-time deployment burdens. These findings help outline the direction for future research by emphasizing larger multilingual corpora, light-weight models efficiently optimized for on-device inference, and learning strategies for multimodal data that generalize across environments and signers.

IV. RESULTS

Aggregated findings from reviewed research indicate a number of consistent trends.

A. Performance Trends

Studies prove:

- Transformer architectures achieve the highest accuracy in large-vocabulary datasets.

- GCN models work stably when the visual features are unreliable.
- Multimodal fusion provides accuracy 5–15% higher than single-input models.

B. Influence of Dataset

Large datasets like WLASL, MS-ASL, and Phoenix-2014T result in remarkable increases in generalization. Smaller regional datasets have lower performances resulting from limited vocabulary and signer diversity.

C. Accuracy Summary

Typical ranges found in literature:

- CNN-based methods: 70–90%
- GCN-based models: 80–92%
- Transformer-based models: 90–96%
- Multimodal approaches: up to 98% for isolated SLR

D. Major Insights

- Transformers are the state-of-the-art currently.
- Multimodal systems provide the best overall robustness.
- Skeleton-based models enhance environmental resilience.
- Dataset size and diversity considerably impact the model's performance.
- Real-time performance remains difficult with high-complexity models.

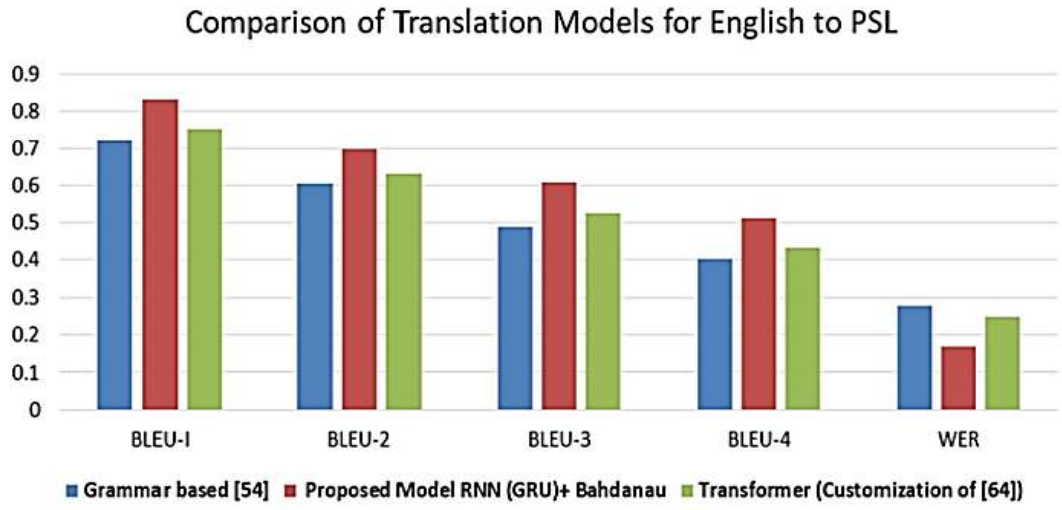

Figure 5: Comparison of transformer and RNN

Figure 5 is a comparative graph between transformer models and RNN models. The graph shows the capabilities of transformers over RNN models in perceiving temporal relationships, thereby increasing accuracy during sign language recognition.

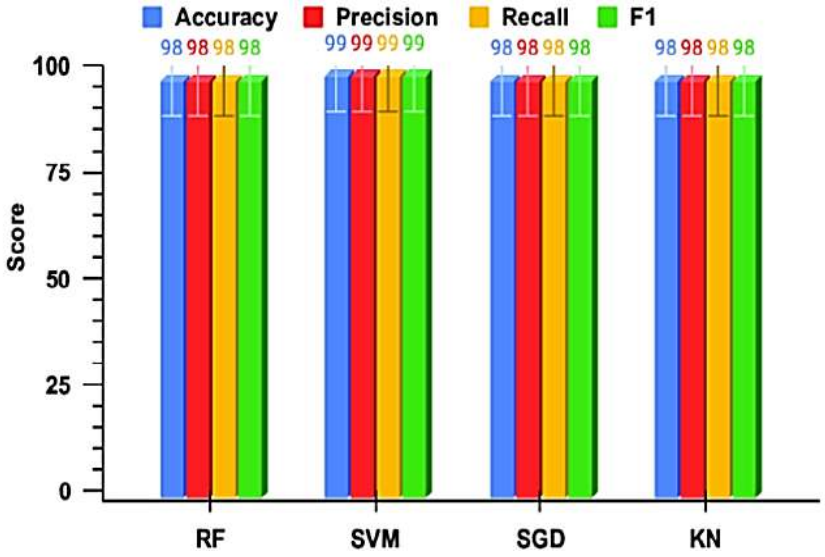

Figure 6: Performance metrics analysis of machine learning techniques with probabilistic features.

In the above Figure 6 shows the performance metrics using probabilistic features for various methods of machine learning. This graph allows for a quantitative analysis of both effectiveness and accuracy of recognition, thus validating the observations made in this section.

V. CONCLUSION

This survey presents an in-depth analysis of major developments in deep-learning-based Sign Language Recognition across various architectures, modalities, and datasets. Recent developments, like transformers and multimodal learning, have substantially increased the accuracy, robustness, and scalability of recognition. Yet, limitations remain, including the limited diversity of datasets, the high inter-signer variability, continuous sign segmentation, and real-time inference with computational demands.

Future work is thus needed in large multilingual data creation, lightweight models for edge deployment, self-supervised and foundation-model learning, signer independence, and integration of techniques for improving real-world applicability through multimodal fusion. With each step, the progress continues to point toward the fact that SLR systems have tremendous potential for transformation into accessibility and communication technologies.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] M. Geetha, N. Aloysius, D. A. Somasundaran, A. Raghunath, and P. Nedungadi, "Toward real-time recognition of continuous Indian sign language: A multi-modal approach using RGB and pose," *IEEE Access*, vol. 13, pp. 60270–60283, 2025. Available from: https://doi.org/10.1109/ACCESS.2025.3554618

[2] M. Al-Qurishi *et al.*, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," *IEEE Access*, vol. 9, pp. 156241–156255, 2021. Available from: https://doi.org/10.1109/ACCESS.2021.3110912

[3] S. M. Miah *et al.*, "Sign language recognition using graph and general deep neural network based on large-scale dataset," *IEEE Access*, vol. 12, pp. 25001–25015, 2024. Available from: https://ieeexplore.ieee.org/document/10456765

[4] Joksimoski *et al.*, "Technological solutions for sign language recognition: A scoping review of research trends, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 101935–101960, 2022. Available from: https://ieeexplore.ieee.org/document/9739689

[5] R. Kothadiya *et al.*, "SIGNFORMER: DeepVision transformer for sign language recognition," *IEEE Access*, vol. 11, pp. 17356–17369, 2023. Available from: https://ieeexplore.ieee.org/document/10011551

[6] Y. Zhang *et al.*, "Sign language recognition based on CNN-BiLSTM using RF signals," *IEEE Sensors Journal*, vol. 24, no. 2, pp. 1456–1467, 2024. Available from: https://ieeexplore.ieee.org/document/10798415

[7] M. Maruyama *et al.*, "Word-level sign language recognition with multi-stream neural networks focusing on local regions and skeletal information," *IEEE Access*, vol. 12, pp. 89012–89026, 2024. Available from: https://ieeexplore.ieee.org/document/10749796

[8] N. Naz *et al.*, "SIGNGRAPH: An efficient and accurate pose-based graph convolution approach toward sign language recognition," *IEEE Access*, vol. 11, pp. 39211–39225, 2023. Available from: https://ieeexplore.ieee.org/document/10049842

[9] Z. Zhou *et al.*, "SignBERT: A BERT-based deep learning framework for continuous sign language recognition," *IEEE Access*, vol. 9, pp. 104895–104907, 2021. Available from: https://ieeexplore.ieee.org/document/9635818

[10] B. A. Al Abdullah *et al.*, "Advancements in sign language recognition: A comprehensive review and future prospects," *IEEE Access*, vol. 12, pp. 74122–74145, 2024. Available from: https://ieeexplore.ieee.org/document/10670380

[11] Li *et al.*, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, 2020. Available from: https://arxiv.org/abs/1910.11006

[12] Li *et al.*, "Transferring cross-domain knowledge for video sign language recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. Available from: https://arxiv.org/abs/2003.03703

[13] N. C. Camgoz *et al.*, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. Available from: https://arxiv.org/abs/2003.13830

[14] M. De Coster *et al.*, "Sign language recognition with transformer networks," in *Proc. International Conf. Language Resources and Evaluation (LREC)*, 2020. Available from: https://aclanthology.org/2020.lrec-1.737

[15] M. Boháček and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," in *Proc. IEEE/CVF WACV Workshops*, 2022. Available from: https://ieeexplore.ieee.org/document/9707552

[16] C. C. de Amorim *et al.*, "Spatial-temporal graph convolutional networks for sign language recognition," in *Proc. Int. Conf. Artificial Neural Networks (ICANN)*, 2019. Available from: https://arxiv.org/abs/1901.11164

[17] J. Huang *et al.*, "Video-based sign language recognition without temporal segmentation," in *Proc. AAAI Conf. Artificial Intelligence*, 2018. Available from: https://arxiv.org/abs/1801.10111

[18] H. Zhou *et al.*, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proc. AAAI Conf. Artificial Intelligence*, 2020. Available from: https://arxiv.org/abs/2002.03187

[19] L. Hu *et al.*, "Continuous sign language recognition with correlation network," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023. Available from: https://arxiv.org/abs/2303.03202

[20] L. Hu *et al.*, "Temporal lift pooling for continuous sign language recognition," in *Proc. European Conf. Computer Vision (ECCV)*, 2022. Available from: https://arxiv.org/abs/2207.08734

[21] L. Hu *et al.*, "Self-emphasizing network for continuous sign language recognition," in *Proc. AAAI Conf. Artificial Intelligence*, 2023. Available from: https://arxiv.org/abs/2211.17081

[22] Hao *et al.*, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021. Available from: https://ieeexplore.ieee.org/document/9710140

[23] Y. Min *et al.*, "Visual alignment constraint for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021. Available from: https://arxiv.org/abs/2104.02330

[24] K.-L. Cheng *et al.*, "Fully convolutional networks for continuous sign language recognition," in *Proc. European Conf. Computer Vision (ECCV)*, 2020. Available from: https://arxiv.org/abs/2007.12402

[25] R. Zuo and B. Mak, "C2SLR: Consistency-enhanced continuous sign language recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*,

2022. Available from: https://ieeexplore.ieee.org/document/9880334
[26] Xiao *et al.*, "SLRFormer: Continuous sign language recognition using vision transformer," in *Proc. ACII Workshops*, 2022. Available from: https://ieeexplore.ieee.org/document/10086026
[27] R. Hinrichs *et al.*, "Continuous sign-language recognition using neural ordinary differential equations," in *Proc. Int. Conf. Pattern Recognition Applications and Methods (ICPRAM)*, 2023.
[28] R. Zuo *et al.*, "Towards online continuous sign language recognition and translation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2024. Available from: https://arxiv.org/abs/2401.05336
[29] N. S. Dinh *et al.*, "Sign language recognition: A large-scale multi-view benchmark and comprehensive evaluation," in *Proc. IEEE/CVF WACV*, 2025. Available from: https://ieeexplore.ieee.org/document/10943527
[30] Tunga *et al.*, "Pose-based sign language recognition using GCN and BERT," in *Proc. IEEE/CVF WACV Workshops*, 2021. Available from: https://ieeexplore.ieee.org/document/9407595
[31] K. M. Dafnis *et al.*, "Bidirectional skeleton-based isolated sign recognition," in *Proc. LREC*, 2022. Available from: https://aclanthology.org/2022.lrec-1.797
[32] D. Li *et al.*, "TSPNet: Hierarchical feature learning for sign language translation," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020. Available from: https://arxiv.org/abs/2010.05468
[33] Sridhar *et al.*, "INCLUDE: A large-scale dataset for Indian sign language," in *Proc. ACM Multimedia*, 2020. Available from: https://dl.acm.org/doi/10.1145/3394171.3413528
[34] S. Naeem *et al.*, "Pakistani word-level sign language recognition based on deep spatiotemporal network," *AAAI Symposium Series*, vol. 6, no. 1, pp. 123–130, 2025. Available from: https://doi.org/10.1609/aaaiss.v6i1.36042
[35] R. Sreemathy *et al.*, "Continuous word-level sign language recognition using an expert system," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 100–110, 2023. Available from: https://doi.org/10.1016/j.ijcce.2023.04.002
[36] J. Bora *et al.*, "Real-time Assamese sign language recognition using deep learning," *Procedia Computer Science*, vol. 215, pp. 785–792, 2023, doi: 10.1016/j.procs.2023.01.117. Available from: https://doi.org/10.1016/j.procs.2023.01.117
[37] B. Baidya *et al.*, "Word-level Nepali sign language recognition using transformer networks," in *Proc. IOE Graduate Conf.*, 2022. Available from: https://conference.ioe.edu.np/publications/ioegc12/IOEGC-12-115-12166.pdf
[38] El Zaar *et al.*, "High performance deep learning applied to multiple sign languages," *E3S Web of Conferences*, vol. 351, 2022. Available from: https://doi.org/10.1051/e3sconf/202235101065
[39] S. Khanna *et al.*, "Sign language interpretation using ensembled deep learning networks," *ITM Web of Conferences*, vol. 53, 2023. Available from: https://doi.org/10.1051/itmconf/20235301003
[40] R. A. Alawwad *et al.*, "Indian regional sign language recognition using CNN models," *Array*, vol. 16, 2022.
[41] Lijiya *et al.*, "SIGNET: A deep learning based Indian sign language recognition system," in *Proc. ICCSP*, 2019. Available from: https://ieeexplore.ieee.org/document/8698006
[42] T. Goswami and S. R. Javaji, "CNN model for American sign language recognition," in *Proc. ICCCE*, 2020. Available from: https://link.springer.com/chapter/10.1007/978-981-15-7961-5_6
[43] M. ElBadawy *et al.*, "Arabic sign language recognition with 3D convolutional neural networks," in *Proc. ICICIS*, 2017. Available from: https://ieeexplore.ieee.org/document/8260028
[44] M. ElBadawy *et al.*, "Deep convolutional neural networks for sign language recognition," in *Proc. SPACES*, 2018. Available from: https://doi.org/10.1109/SPACES.2018.8316344
[45] L. Pigou *et al.*, "Sign language recognition using convolutional neural networks," in *Proc. ECCV Workshops*, 2014. Available from: https://link.springer.com/chapter/10.1007/978-3-319-16178-5_40
[46] L. Pigou *et al.*, "Gesture and sign language recognition with temporal residual networks," in *Proc. ICCV Workshops*, 2017. Available from: https://tinyurl.com/mu4ww2hy
[47] Hu *et al.*, "Hand-model-aware sign language recognition," in *Proc. AAAI Conf. Artificial Intelligence*, 2021. Available from: https://doi.org/10.1609/aaai.v35i2.16247
[48] Hu *et al.*, "SignBERT: Pre-training hand-model-aware representation for sign language recognition," in *Proc. IEEE/CVF ICCV*, 2021. Available from: https://tinyurl.com/45y2hvez
[49] C. Lu *et al.*, "Sign language recognition with multimodal sensors and deep learning methods," *Electronics*, vol. 12, no. 23, 2023. Available from: https://doi.org/10.3390/electronics12234827
[50] R. Rastgoo *et al.*, "Sign language recognition: A deep survey," *Expert Systems with Applications*, vol. 164, 2021. Available from: https://doi.org/10.1016/j.eswa.2020.113794