Department of Computer Science and Engineering

T.K.M College of Engineering, Kollam

NOVEMBER 2024

SIGN LANGUAGE RECOGNITION USING GRAPH AND DEEP NEURAL NETWORK BASED ON LARGE SCALE DATASET

Seminar Report

Submitted By

BHAGYA A JAI (TKM21CS041)

to

APJ Abdul Kalam Technological University in partial fulfilment of the requirements for the award of B.Tech Degree in Computer Science and Engineering

DECLARATION

I undersigned hereby declare that the report on “Sign Language Recognition Using Graph and General Deep Neural Network Based on Large Scale Dataset”, submitted as part of the course, Seminar, under APJ Abdul Kalam University, Kerala is a bonafide work done by me under the supervision of Dr. Aneesh G. Nath, Head of Department, Department of Computer Science and Engineering, and Dr. Ansamma John, Professor, Department of Computer Science and Engineering, TKM College of Engineering.

This submission represents my ideas in my own words and from other sources that have been adequately and accurately cited and referenced. I also declare that I have adhered to the ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission.

I understand that any violation of the above will be a cause for disciplinary action by the Institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other
University.

Ms. Bhagya A Jai

Place: Kollam

Date: 05/11/2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING TKM COLLEGE OF ENGINEERING, KOLLAM

CERTIFICATE

This is to certify that the report titled "**Sign Language Recognition Using Graph and General Deep Neural Network Based on Large Scale Dataset**" submitted by **Bhagya A Jai, TKM21CS041** to the APJ Abdul Kalam Technological University in completion of the requirements for the award of Bachelor of Technology Degree in Computer Science and Engineering during 2024 – 2025 is a bonafide record of the **Seminar** carried out by her under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Prof. Ansamma John
Seminar Coordinator
Professor
Dept.of CSE
TKM College of Engineering
Kollam

Dr. Aneesh G Nath
Head of Department
Associate Professor
Dept.of CSE
TKM College of Engineering
Kollam

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude to the Almighty and sincere thanks to all who helped me to complete the seminar successfully.

I express my sincere gratitude to **Dr. Sajeeb R,** Principal, TKMCE, for providing me with all the necessary facilities and support for doing the seminar.

I am extremely grateful to **Dr. Aneesh G. Nath**, Head of Department, Department of Computer Science and Engineering, **Dr. Ansamma John**, Seminar Coordinator, and Professor, Department of Computer Science and Engineering, and **Dr. Manu J Pillai**, Associate Professor, Department of Computer Science and Engineering, for their constructive guidance, advice, constant support and technical guidance provided throughout the preparation of this seminar. Without their intellectual support and appropriate suggestions at the perfect time, this seminar would not have been possible.

I extend my immense gratitude to all faculties and technical staff in the Department of Computer Science and Engineering, for their help and necessary facilities to complete the seminar. My humble gratitude and heartiest thanks also go to my parents and friends, who have supported and helped me on the course of this work.

Ms. Bhagya A Jai

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

Abbreviations	**Full Form**
GCAR	Graph Convolution with Attention and Residual
GC-AM	Graph Channel Attention Module
Sep-TCN	Separable Temporal Convolutional Network
WLASL	Word-Level American Sign Language Dataset
MSL	Mexican Sign Language Dataset
ASLLVD	American Sign Language Lexicon Video Dataset

ABSTRACT

Sign Language Recognition (SLR) is a transformative technology that bridges communication between hearing-impaired and non-hearing-impaired communities, moving beyond traditional interpreter-based methods. Current SLR methods often fall short due to challenges in capturing complex gestures, spatial nuances, and temporal consistency, which are essential to accurately interpret sign language. Many existing SLR systems focus primarily on hand skeleton data to address issues like partial occlusion and background noise, yet they frequently overlook the role of whole-body motion, facial expressions, and contextual environmental factors. The proposed two-stream multistage Graph Convolution with Attention and Residual Connection (GCAR) model offers a comprehensive solution by capturing extensive spatial-temporal contextual information. In the GCAR model, joint key features and joint motion data are processed through two parallel streams. The first stream employs Separable Temporal Convolutional Networks (Sep-TCN), Graph Convolution layers, and a Channel Attention Module to extract spatial-temporal features from static joint data, while the second stream handles dynamic joint motion, generating complementary features. The dual-stream setup allows the model to retain high-level information on joint movements and expression details, which are vital for distinguishing between subtle variations in signs. These streams are then fused to create a robust final feature vector, optimized for classification accuracy through multi-stage attention layers and residual connections that enhance feature retention and prevent degradation. Despite its depth, the model is computationally efficient, with only 0.69 million parameters, making it suitable for real-time applications. Extensive tests on large-scale datasets such as WLASL, PSL, MSL, and ASLLVD demonstrate the model's exceptional performance, achieving accuracy rates of 90.31%, 94.10%, 99.75%, and 34.41%, respectively, surpassing previous benchmarks. Furthermore, the GCAR model introduces a novel feature fusion technique with a dynamic attention mechanism that selectively focuses on essential spatio-temporal cues, and boosting recognition accuracy across varied lighting and backgrounds. The approach provides an accessible communication platform for the hearing-impaired and represents a major step forward in SLR technology by not only enhancing classification accuracy but also improving generalizability, making it adaptable across diverse sign languages.

1. INTRODUCTION

In the modern world, communication is paramount, yet for the hearing-impaired community, effective communication can be a significant challenge. Unlike spoken languages, sign language is visual and involves specific hand gestures, body language, and facial expressions to convey meaning. While many spoken languages are widely recognized and taught, sign language remains largely unique to specific regions or cultures. This linguistic diversity and complexity add challenges for hearing-impaired individuals who seek to interact seamlessly with hearing individuals in society. Recent statistics reveal that approximately 5% of the world's population, which equates to about 466 million individuals, including both adults and children, is hearing impaired. Given these statistics, there is an urgent need for accessible methods to bridge this communication gap. In this context, the advancement of automated sign language recognition systems becomes an essential tool to enable the hearing-impaired community to communicate effectively with others, facilitating their access to education, healthcare, and employment.

1.1. Sign Language Recognition: Unique Structure and Challenges

Sign language serves as a unique linguistic structure, differing markedly from spoken languages like English or Hindi, as it relies on spatial signs, body movements, facial expressions, and gestures. This multifaceted communication style allows for conveying nuanced emotions and intentions but introduces challenges for both learners within the hearing-impaired community and non-signers. The lack of formal training resources makes mastering sign language difficult for many hearing-impaired individuals, while the general public often does not learn it, leaving communication gaps. Sign language is also region-specific; for instance, American Sign Language (ASL) and British Sign Language (BSL) vary despite the shared spoken language of English.

Developing robust Sign Language Recognition (SLR) systems also faces technical hurdles, particularly with dynamic, real-time interpretation. Early models were limited to static signs, which convey individual meanings, whereas real-world communication involves dynamic signs that consist of gesture sequences. Capturing these gestures accurately is challenging, given variability in expression and high intraclass variability among users. Systems focused solely on hand movements often misinterpret signs as they

miss the whole-body cues, facial expressions, and emotional context. Addressing these gaps requires SLR systems that incorporate facial, whole-body, and contextual environmental data to provide a holistic understanding of signs, enhancing communication accuracy and effectiveness.

1.2. Advancements in Recognition Systems and the Proposed Model

Modern SLR research leverages advanced sensor-based and vision-based approaches to capture sign language nuances. Sensor-based methods, though highly accurate, face challenges related to portability and cost. In contrast, vision-based methods use cameras to capture gestures and have gained traction for their adaptability and cost-effectiveness. Among vision-based approaches, skeleton-based data—recording 2D or 3D coordinates of joints—has become popular, especially with advancements in pose detection technologies like MediaPipe and OpenPose. However, conventional skeleton-based models often depend on handcrafted features or static spatial data, making it difficult to recognize the temporal dynamics of dynamic gestures. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks may miss intricate relationships in joint movements.

To overcome these limitations, this study introduces a two-stream multistage Graph Convolution with Attention and Residual (GCAR) model, designed to capture spatial-temporal relationships in dynamic signs. The model operates two parallel streams: one for joint skeletal information and the other for motion data, both of which apply Separable Temporal Convolutional Networks (Sep-TCN). These streams are then combined using attention modules, enhancing discriminative capability by focusing on non-connected skeletal details. A Channel Attention Module in the GCAR framework dynamically adjusts focus based on motion and structure, which significantly enhances classification accuracy and model generalizability. Tested on large-scale datasets like ASL and PSL, the GCAR model has shown notable improvements in performance, handling both static and dynamic signs effectively and establishing new benchmarks in SLR. Through these advancements, this work aims to enhance inclusivity and accessibility for the hearing-impaired community, setting a new standard in automated sign language recognition technology.

2. LITERATURE SURVEY

Extensive research has been conducted in the field of sign language recognition (SLR), exploring various models, techniques, and challenges. Recently, some researchers have proposed frameworks that attempt to establish a parallel between sign structures and phonological elements in spoken languageoe paradigm, which defines sign language components through shape, alignment, and motion, is often used to categorize signs . Key aspn language, such as hand orientation, layout of fingers, and static versus dynamic postures, carry distinct meanings. Static gestures convey complete meaning with a single frame, while dynamic gestures require sequential frames to convey context .

Two primary apprSLR include vision-based systems, favored for their accessibility and low cost, and sensor-based systems. Vision-based SLR employs diverse statistical and machine learning techniques, and recent work has made strides with deep learning models. For example, Raghuveera et al. used Histogram of Oriented Gradients (HOG) for feature extraction from RGB images and achieved a 78.85% accuracy on the ISL dataset using Support Vector Machines (SVM) while Solis et al. utilizedrier Moments (JFMs) with an Artificial Neural Network (ANN) to achieve 95% accuracy on the MSL dataset .With large-scale datasets, machine levels showed limitations in scalability and accuracy, leading to an increased focus on deep learning approaches. Miah et al. demonstrated better accuracy with Convolutional Neural Networks (CNNs) on a segmented ASL dataset , and Neverova et al. proposed a multi-modality using RGB images and audio, though the audio component was less applicable for SLR .

In terms of sequential recognition, Put et al. combinedist Temporal Classification (CTC) and 3D Convolutional Residual Networks (3D-ResNet) for better feature learning and achieved reduced word error rates (WER) on CSL and RWTH-PHOENIX-Weather datasets . Koller et al. applied a CNN-Hidden Markov Model (HMM) hybrid, performance improvements across multiple datasets . Other innovative methods include Huang et al.'s use of a 3D CNN with an attention for spatial-temporal feature extraction, which reported high accuracy on the ChaLearn and CSL datasets .For feature extraction in video-based datasets, researchers have explored various methods. Pigou et al. proposed a temporal pooling method , while Sincan et al. combined Long Short-Term Memory (LSTM) and CNN with VGG16 for an Italian sign dataset, achieving an accuracy of 93.15% . Advanced techniques like the Hierarchical Attention Network (HAN) by Huan et al. have also been

employed segmentation in continuous SLR, though their approach is limited to pure visual features.

A major limitation in existing SLR models is their struggle with background noise, hand occlusion, illumination issues. To address these, some researchers have shifted to skeleton-based SLR, using tools like OpenPose and MediaPipe for skeletal point extraction, focusing on spatial data to enhance gesture recognition. For example, Musa et al. utilized a Graph Convolutional Network (GCN) with attention mechanisms across skeleton points, while Yan et al.ed spatio-temporal GCNs with hand joint extraction .To further enhance inter-gesture variance, recent approaches integrate full-body and facial information. Solis et al., for instance, used body, ns from a spatial camera with LSTM and RNN to achieve improved performance in MSL recognition. Xia et al. and Peraz et al. similarly achieved high accuracy using datasets with comprehensive body key points, though these models often lack scalability due to scalloped datasets. The static-based recognition systems in these models hinder generalization, especially for dynamic signs.

Li et al. addressed the limitations of smaller datasets by developing a dataset with 2000 classes. Their application of Temporal Graph Convolutional Networks (TGCN) achieved a 62.63% top-10 accuracy, highlighting the value of extensive datasets and temporal features for SLR .

In this research, a multistage graph convolutional network with attention and residual connections (GCAR) is proposed to enhance the accuracy and efficacy of SLR systems, aiming to address these challenges.

3. DATASETS

For the evaluation of the proposed sign language recognition model, a selection of benchmark datasets was employed, including the Word-Level American Sign Language (WLASL) dataset, Malaysian Sign Language (MSL) dataset, American Sign Language Lexicon Video Dataset (ASLLVD), and Pakistani Sign Language (PSL) dataset. Each of these datasets brings unique characteristics in terms of the specific sign language represented, dataset scale, and variations in data collection methodologies, such as the number of signers, video quality, and environmental factors. These distinctions provide a robust foundation for assessing model performance across a range of sign languages and testing the model's ability to generalize effectively.

In the following sections, each dataset is discussed in greater detail, highlighting its individual attributes and how it contributes to a comprehensive and diverse evaluation of the model's capabilities.

TABLE 1. Evaluated dataset description in the study

Dataset Names	Lang.	Signs	Sub.	Total videos	Videos Per Sign	Joint Per Frame
WLASL	ASL	2000	119	21089	10.5	67
MSL	MSL	30	20	3000	20	67
ASLLVD	ASL	2745	n/a	9748	3/4	67
PSL	PSL	19	n/a	2700 (img)	55 (img)	67

TABLE 2. The description of the 67 whole body key points

Sr No	Hand Pose No	L. Hand Pose Name	Sr No	Pose No	Hand Pose Name	Sr No	Hand Pose No	Righ Hand Pose Name
1	0	Wrist	20	19	Ring MCP	36	9	Middle CMC
2	1	Thumb CMC	21	20	Ring IP	37	10	Middle MCP
3	2	Thumb MCP		Body Pose	Body Pose Name	38	11	Middle IP
4	3	Thumb IP	22	0	Nose	39	12	Middle TIP
5	4	Thumb TIP	23	11	Lef shoulder	40	13	Ring CMC
6	5	Index CMC	24	13	Left elbow	41	14	Ring MCP
7	6	Index MCP	25	12	Right shoulder	42	15	Ring IP
8	7	Index IP	26	14	Left elbow	43	16	Ring TIP
9	8	Index TIP		Hand Pose No	Right Hand pose	44	17	Right heel
10	9	Middle CMC	27	0	Wrist	45	18	Ring CMC
11	10	Middle MCP	28	1	Thumb CMC	46	19	Ring MCP
12	11	Middle IP	29	2	Thumb MCP	47	20	Ring IP
13	12	Middle TIP	30	3	Thumb IP		Face	Facial Landmark
14	13	Ring CMC	31	4	Thumb TIP	48-51	4	Right eyebrow
15	14	Ring MCP	32	5	Index CMC	52-55	4	Left eyebrow
16	15	Ring IP	33	6	Index MCP	56-59	4	Right eyebrow
17	16	Ring TIP	34	7	Index IP	60-63	4	Left eyebrow
18	17	Right heel	35	8	Index TIP	64-67	4	Mouth
19	18	Ring CMC						

3.1. WLASL Dataset

The Word-Level American Sign Language (WLASL) dataset is one of the most extensive collections for ASL, comprising 68,129 videos with 20,863 unique ASL glosses from 20 sources. Each video features a signer performing a single sign, recorded primarily from a frontal view but with varied backgrounds to increase complexity. Preprocessing excluded multi-word gloss annotations, retaining only single-word labels. Glosses were organized by frequency, resulting in subsets of different vocabulary sizes—WLASL100, WLASL300, WLASL1000, and WLASL2000—allowing an examination of model scalability and the challenges of word-level sign recognition across varying vocabulary ranges.

3.2. MSL Dataset

The Mexican Sign Language (MSL) dataset contains 30 distinct signs, each recorded using an OAK-D camera that captures 67 key points across both hands, the body, and the face, utilizing the MediaPipe library. Each sign was recorded 25 times, totaling 3000 samples and extracting 20 frames per video, thus ensuring robust sample coverage for high-precision recognition of Malayalam signs.

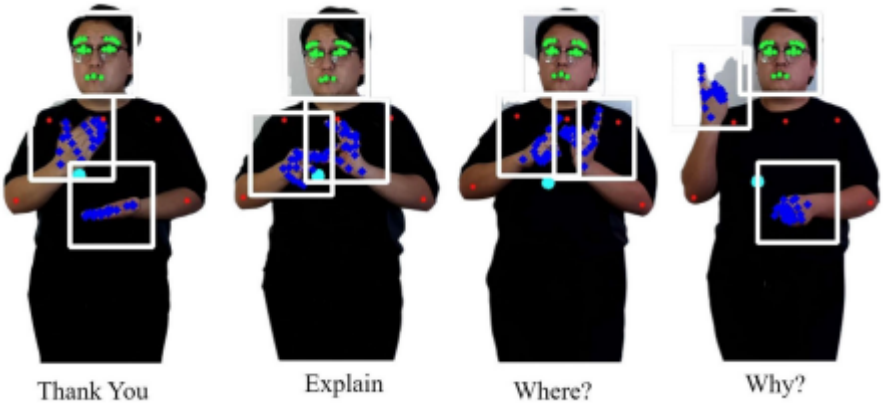

FIGURE 2. Sample images extracted from the Mexican Sign Language (MSL) dataset

3.3. ASLLVD Dataset

The American Sign Language Lexicon Video Dataset (ASLLVD) offers a large-scale collection of 2745 unique ASL words, recorded from multiple angles to enable synchronized, multi-view analysis. This dataset includes 7798 training videos and 1950 test videos, each annotated with gloss labels, start and end frames, and detailed morphological classifications such as handshape and articulation. The multi-view and annotated nature of ASLLVD enhances the model's ability to learn from varied perspectives, with unique ID labels and raw video sequences supporting the accuracy of recognition tasks.

FIGURE 3. Sample images from American Sign Language Lexicon Video Dataset (ASLLVD).

3.4. PSL Dataset

The Pakistani Sign Language (PSL) dataset includes skeleton-based body, face, and hand key points recorded from nine participants. Data were collected using a standard web camera, the MediaPipe and OpenPose libraries, and saved in JSON format. Each frame contains 67 key points (21 per hand and 25 for the body and face). This dataset features both alphabet and word signs (12 word signs and 37 alphabet signs) and was standardized post-recording to normalize hand and body positions across signers.

Each dataset contributes unique strengths in terms of sign language specificity, recording perspectives, and annotation methods, providing a comprehensive benchmark for evaluating the model's adaptability and performance across multilingual sign language applications.

4. PROPOSED METHODOLOGY

In the study, we propose a robust two-stream model designed to extract spatial-temporal contextual information for sign language recognition. The model leverages multistage Graph Convolution with Attention and Residual connection (GCAR), enabling dynamic attention across non-connected skeleton points during key events. With input sequences represented as $X(1:N) = [x_1,x2,x3,\ldots,xN]$ over N frames, each pose frame is a concatenated vector of skeleton key points with dimension K. The methodology involves a two-stream model, where joint skeleton points and joint motion are processed to capture complete body movements in sign language gestures.

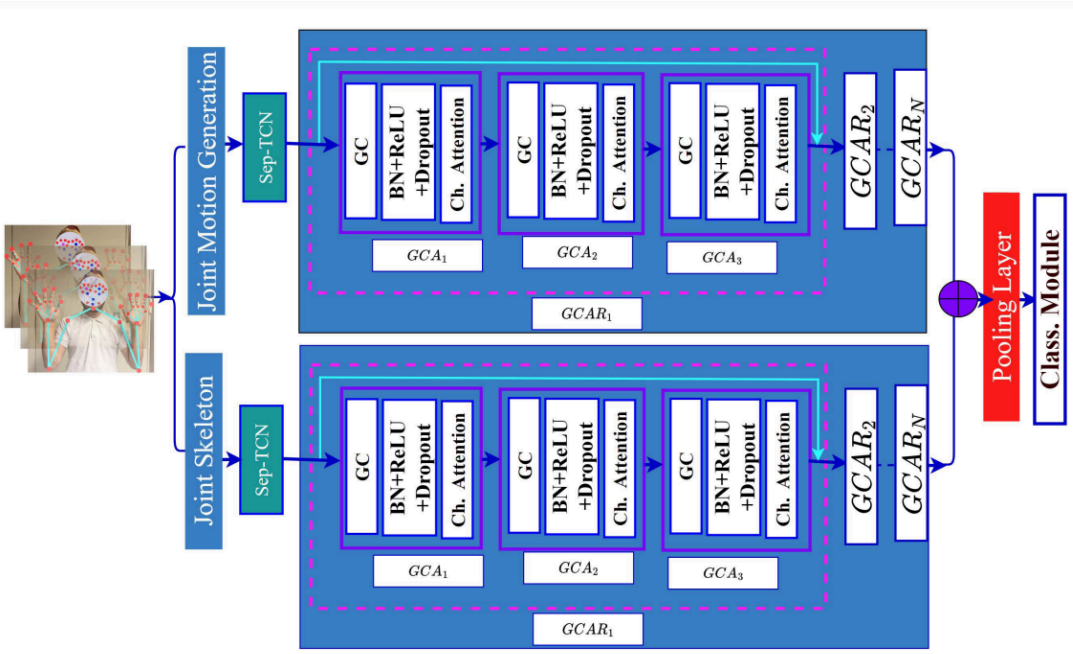

FIGURE 4. Working flow architecture.

The proposed model consists of two primary streams:

First Stream: Processes joint key points with Sep-TCN, GCN, and deep learning modules, concluding with a Channel Attention (CA) module to focus on channel-wise effective features.

Second Stream: Processes joint motion similarly, yielding features combined with the first stream's output for final classification.

This GCAR approach capitalizes on Channel Attention and Separable TCN (Sep-TCN) for high-performance accuracy, handling complex spatial-temporal dependencies with 22 stages in both streams.

4.1. Pose Estimation

In this phase, skeleton information is extracted, leveraging hand gesture data for enhanced privacy and security. Using a dataset of 67 key points (42 from hands, 5 from the body, and 20 from the face), the system identifies skeletal structures and gestures without exposing sensitive appearance details. By protecting biometric data such as palm prints and fingerprints, the model allows secure hand gesture recognition while preserving individual privacy. Face, body, and hand data integration provides a comprehensive view of gestures, as illustrated in Figure 1, which visualizes the 67 extracted key points.

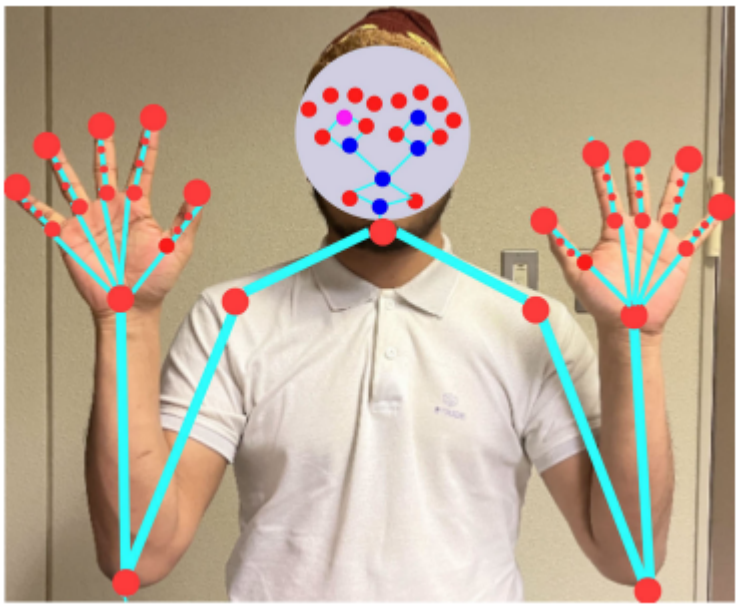

FIGURE 1. Pose and graph construction.

4.2. Motion Calculation

Dynamic signs play a crucial role in various sign language word recognition systems. To compute effective features for dynamic sign language, motion is extracted from the input frame sequence prior to feeding into the graph convolution (GC) model. This emphasizes the importance of movement and alignment in enhancing the efficiency of the skeleton-based data structure.

In developing the dynamic sign language recognition system, joint skeleton movement is considered a critical feature influenced directly by motion. We compute a 2-dimensional vector based on the x and y coordinates from the 67 skeleton key points, evaluating the joint position differences between consecutive frames. The motion calculation is visualized in Figure 5, and it can be computed using the following equations:

$$\text{Motion}_X = X_t - X_{t+1}$$

$$\text{Motion}_Y = Y_t - Y_{t+1}$$

$$\text{Motion}_Z = Z_t - Z_{t+1}$$

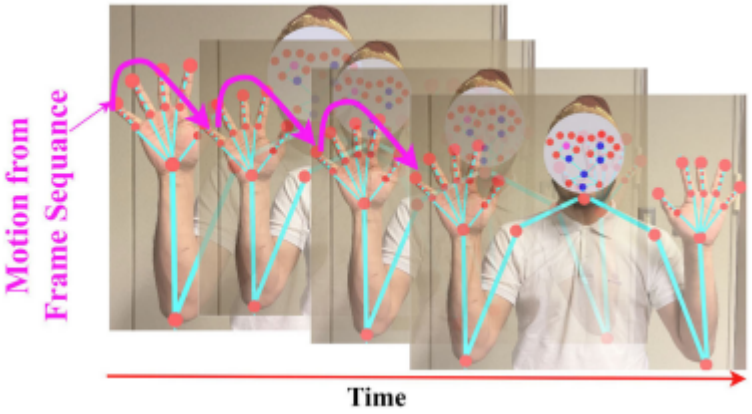

FIGURE 5. Motion calculation demonstration

4.3.Separable TCN

Separable TCN, a specialized variant of Temporal Convolutional Networks (TCN), is employed to optimize computational efficiency while maintaining robust model performance. Unlike traditional TCNs, Separable TCN decomposes the convolution process into two distinct stages: depth-wise (DW) and pointwise (PW) convolution. This strategic division minimizes the number of required parameters and significantly reduces computation costs, represented in Floating Point Operations Per Second (FLOPS).

The mathematical operations for each time in depth-wise convolution require K×K×1, while pointwise convolution necessitates 1×1×C filter operations, resulting in a notably lower number of operations per kernel during convolutions. In this study, we apply a 3×1 filter with a stride of one for the first Sep-TCN layer and a 5×1 filter with a stride of two for the second Sep-TCN layer, leading to filter sizes of 13×1×1 and 11×1×1 for depth-wise operations.

Figure 6(a) depicts the internal structure of the Separable TCN, which includes depth-wise and pointwise convolutions, residual connections, and max-pooling layers to enhance temporal features. This configuration focuses on extracting the dominant frame as a feature vector, as demonstrated in Figure 6(b), which elucidates the local temporal proximity extraction across a broader scope.

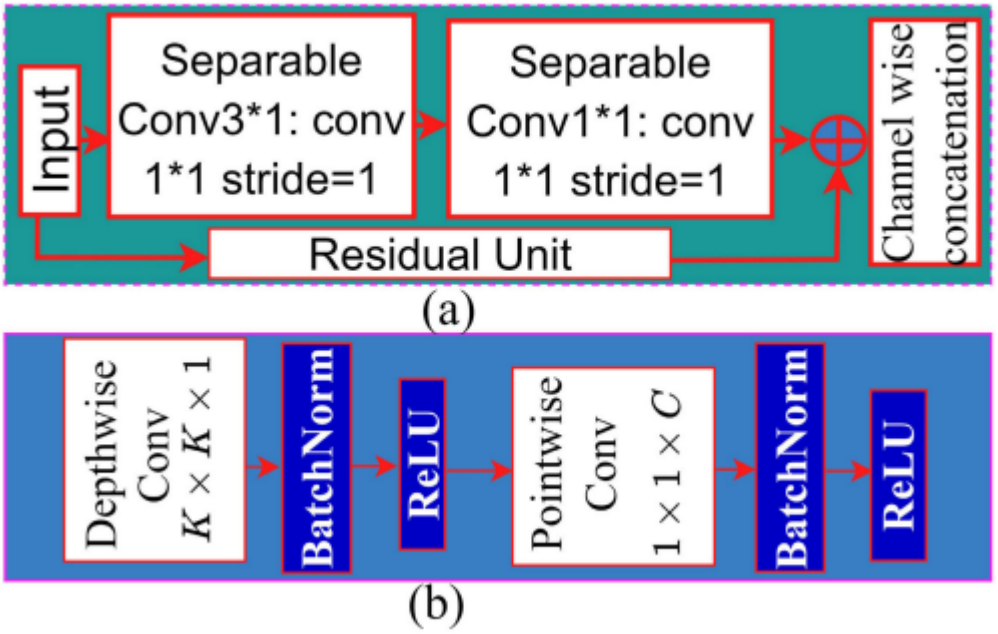

FIGURE 6. (a) SepTCN and (b) Internal structure of the separable TC.

4.4. Graph convolution with attention (GCA) module

The GCA module processes the motion features to extract both local and global range features pertinent to spatial and temporal contextual information. Within the GCA module, a Graph Convolution layer is utilized to enhance long-range dependencies, while a series of deep learning layers—including batch normalization, ReLU activation, and dropout layers—are employed to boost short-range dependencies. Additionally, a channel attention layer enhances channel-wise effective features, allowing the system to efficiently manage non-connected skeleton points and achieve high-performance accuracy.

4.4.1. Graph construction

The dynamic datasets employed for sign language gestures present sequences of frames, where each frame generates a joint skeleton sequence of vectors. In this study, each joint skeleton produces a 2D vector, which

comprises dynamic sign language gestures characterized by varying numbers of frames. To establish a spatial-temporal graph, the joint skeleton information is utilized, with each motion serving as a graph node, and natural connections among joints constituting the edges.

Figure 1 illustrates an example of graph construction focused on the upper body segment. The primary objective is to generate higher-level feature maps on the graph through multiple stages of spatial-temporal graph convolution with attention (GCA) operations. The construction of the graph initiates with a fully connected graph containing N nodes. The edges of this graph are formed based on a weighted adjacency matrix as per the following equation:

$$G = (V, E)$$

In this framework, V and E denote the vertices and edges of the graph. In our study, the 67 key points are considered vertices, expressed as:

$$V = \{v_{(t,i)} \mid t = 1, \ldots, T, i = 1, \ldots, N\}$$

where N=67. Furthermore, the dataset comprises 32 frames from each video.

The ultimate feature is calculated through two processes: (i) capturing the intraframe joint skeleton connection, which primarily explores natural connections among human joints, and (ii) capturing the interframe consecutive joint skeleton connection. These processes delineate spatial and temporal contextual information. The edges of the graph are computed using the weighted adjacency matrix, defined as follows:

$$A \in R^{NxN}$$

Despite the natural human body exhibiting partial connectivity—where hand, palm, and face landmarks are connected, but with no direct connection between the face and hand skeletons—a fully connected graph is constructed to capture joint dependencies through graph learning.

4.4.2. Graph convolution

Graph Convolution is a neural network framework specifically designed to handle graph-structured data, such as the human body's joint skeleton, where joints are nodes and their connections are edges. Unlike traditional CNNs that work with grid data like images, Graph Convolution Networks (GCNs) aggregate information from neighboring nodes to update the central node's features, effectively capturing complex relationships.

In our system, we process a sequence of frames at a specific time t. The number of joint skeleton nodes in a frame is denoted as V^t, and edges are defined as $Edge_t=\{v_i,v_j \mid T=t, (i,j) \in H\}$, where H represents the fully connected human body graph. This setup is crucial for understanding gestures in sign language.

The general convolution equation is:

$$F_{(out)}(x) = \sum_{h=1}^{K} \sum_{w=1}^{K} data_{in}(S(x,h,w)) \cdot W(h,w)$$

For graph data, it is expressed as:

$$F_{out}(v_i) = \sum_{(v_j) \in B_i)} \frac{1}{Z_{ij}} data_{in}(v_j) \cdot w(l_i(v_j))$$

Here, Z normalizes contributions from nodes, and w(·) generates a weighted matrix to assign importance based on the input.

The Graph Convolution operation is defined as:

$$data_{n+1} = GC_n(data_n)\sigma(A_n \cdot data_n \cdot W_n)$$

In this equation, A is the adjacency matrix, $data_n$ is the input, and W is the weight matrix. The activation function σ introduces non-linearity, enhancing learning capabilities.

By employing Graph Convolution, our system effectively processes dynamic sign language data, capturing essential spatial and temporal dependencies among joints for improved gesture recognition.

4.4.3. Channel attention

The Channel Attention model enhances the channel-wise features within the Graph Convolutional Attention Residual (GCAR) framework by emphasizing important features and suppressing less significant ones. This process improves the model's generalization and discriminative ability, as illustrated in Figure 4.

In the GCAR, motion features are processed alongside structured data, creating a feature matrix with multiple channels. The channel attention mechanism generates a weight vector for each channel, allowing the model to dynamically adjust feature importance. It begins with global average pooling, followed by two fully connected layers and batch normalization to mitigate internal covariate shift.

ReLU activation functions are applied to ensure non-negative values, and the final output is obtained by multiplying these activations with the original GCAR features. This operation preserves significant features while diminishing the influence of less important ones. The overall architecture, shown in Figure 7, highlights the sequential operations that contribute to effective feature refinement and selection.

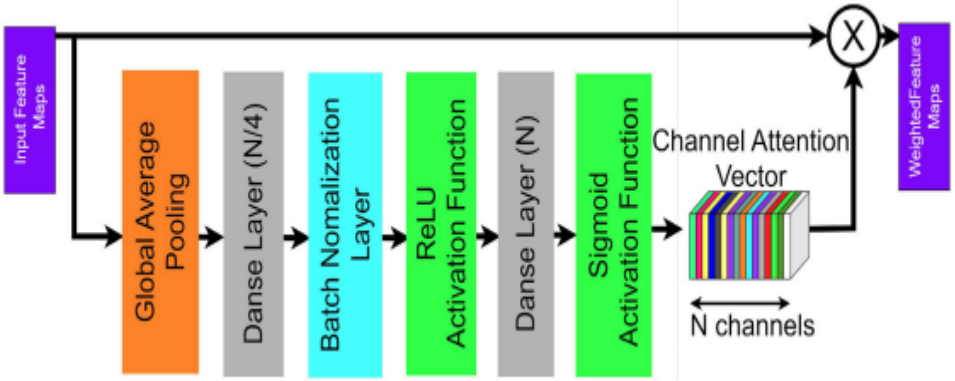

FIGURE 7. Channel attention module.

4.5. Multi-Stage Graph Convolution with Attention and Residual Connection

The GCAR approach integrates multiple Graph Convolution layers with attention mechanisms and residual connections, as illustrated in Figure 4. It consists of three

stacked Graph Convolutions, each incorporating a Channel Attention (GCA) module. The GCAR processes motion features by applying convolution on the initial feature map with dimensions N×T×CN \times T \times CN×T×C, where NNN represents vertices, TTT indicates temporal length, and CCC denotes the number of channels.

The architecture of the GCA network includes a Graph Convolutional layer, ReLU activation, batch normalization, a dropout layer, and the channel attention module. Each of the three GCA sequences outputs feature representations, which are then concatenated with a residual connection. In total, 22 GCAR modules are employed sequentially to form a multi-stage architecture, enhancing the model's ability to capture complex patterns over time, as depicted in Figure 8.

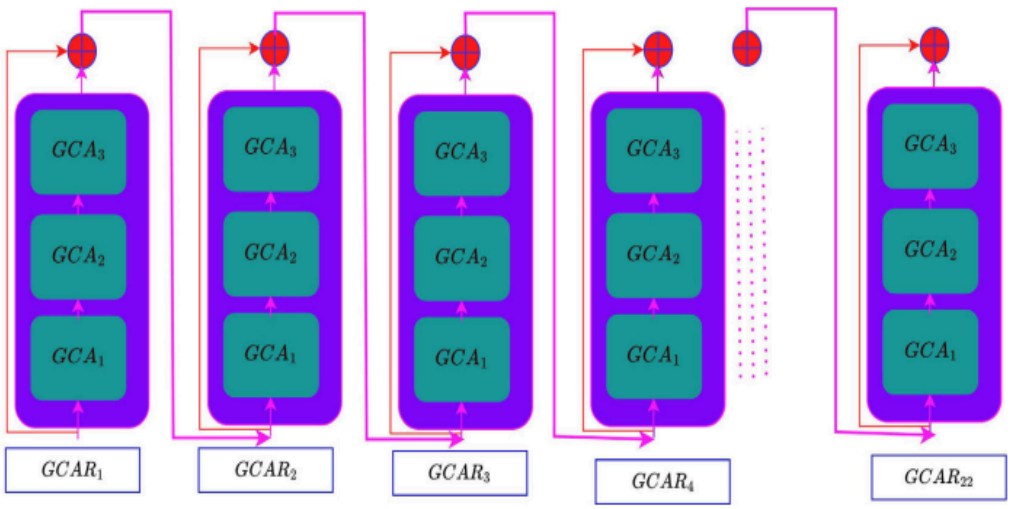

FIGURE 8. Stages of the GCAR model.

4.6. Classification Module

The classification module receives the final feature produced by the GCAR, focusing on refining the features and improving classification performance. It includes several key elements: paired ReLU activation functions, a fully connected layer, batch normalization to standardize the data, a dropout layer to reduce overfitting, and averaging to transform the matrix into a feature vector. The architecture of this module, depicted in Figure 9, demonstrates its organized design for effectively processing and refining features for accurate classification

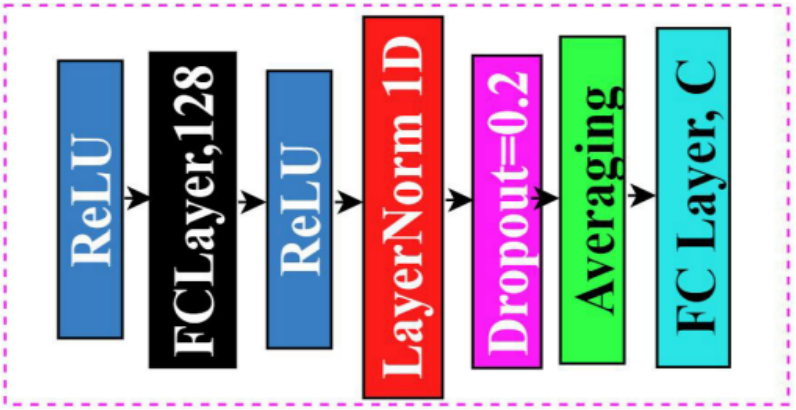

FIGURE 9. Classification module.

5. EVALUATION AND PERFORMANCE

This section outlines the evaluation experiments conducted to validate the proposed model's effectiveness using four benchmark datasets. The experimental setup, evaluation metrics, and performance accuracy are discussed, along with comparisons to state-of-the-art methods.

5.1. Experimental setting

The gloss samples were split into training and testing sets in a 4:2 ratio, ensuring each split contained at least one sample per gloss. This strategy mirrors the method used in prior state-of-the-art approaches. For the other datasets, 70% was allocated for training, and 30% for testing. The model was implemented on a machine with a GeForce RTX 3090 GPU (24GB), 32GB RAM, CUDA version 11.7, and NVIDIA driver version 515. The learning rate was set at 0.003, and the batch size was 32. The final layer's output neurons matched the number of classes in the datasets, such as WLASL 2000 and ASLLVD 2745. The system ran for 300 epochs using the Adam optimizer. Python packages used included PyTorch, alongside OpenCV, pandas, and other processing tools.

5.2. Evaluation metric

The model assessment involved calculating mean scores for top-K classification accuracy (K = 1, 5, 10), accounting for all sign instances. Figure 2 shows that similar meanings can lead to classification errors, although contextual information can mitigate these misclassifications. Hence, using top-K predicted labels is a more practical approach for word-level sign language recognition.

5.3. Ablation study

The ablation study in Table 3 examines how attention mechanisms affect model performance across different configurations of the WLASL dataset, specifically WLASL-100, WLASL-300, and WLASL-2000. It compares performance metrics for Top-1, Top-5, and Top-10 accuracy. The best accuracy of 56.50% was achieved for WLASL-100, outperforming models without attention. Notably, WLASL-300 reached a Top-5 accuracy of 38.47% with the attention module. The configurations

without attention served as baselines, while the inclusion of attention improved accuracy across all datasets, indicating its importance in gesture recognition.

TABLE 3. Ablation study of the proposed model for the WLASL dataset

Dataset Class	Top-1 Accuracy		
	Without Attention (55 Joints)	Without Attention (67 Joints)	Including Attention (67 Joints)
WLASL -100	55.43	56.43	56.50
WLASL-300	38.32	38.00	38.47
WLASL-2000	23.65	23.00	24.10

5.4. Performance accuracy and state of the art comparison of the WLASL dataset

Table 4 presents the performance of both single-stream and two-stream models, detailing their effectiveness across various datasets. Metrics like Top-1, Top-3, Top-5 indicate the accuracy of the model's predictions among the classes, highlighting the advantages of dual data streams in enhancing performance.

TABLE 4. Performance accuracy with WLASL for various configuration

Dataset Class	Top-1		Top-3		Top-5	
	Single Stream	Two Stream	Single Stream	Two Stream	Single Stream	Two Stream
WLASL -100	56.50	63.25	70.00	71.82	80.00	83.33
WLASL-300	38.47	43.80	56.00	57.00	69.01	69.31
WLASL-2000	24.10	24.10	44.00	45.00	52.62	52.62

5.5. Performance accuracy and state of the art comparison for the PSL dataset

Table 6 analyzes the proposed Two-Stream model against its Single Stream counterpart and a state-of-the-art dynamic graph-based model. The Two-Stream model achieved an accuracy of 94.01%, while the Dynamic GCN method achieved 90.00%, and the Single Stream reached 92.40%. Additionally, Figure 10 shows that

the Two-Stream configuration exceeded 85.00% accuracy for most labels, demonstrating its capability in handling the complexity of sign language.

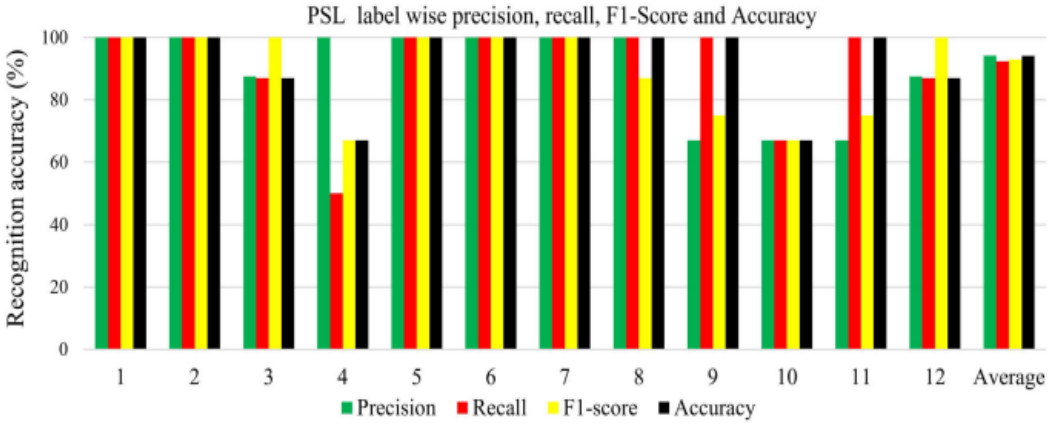

FIGURE 10. Label wise precision, recall and F1-score for the PSL alphabet dataset.

TABLE 5. Performance accuracy and comparison for the PSL dataset.

Method Name	Dataset	Joints Key points	Performance Accuracy [%]
Dynamic-GCN [14]	PSL	Full Body	90.00
Proposed Model (Single Seram)	PSL	Full Body	92.40
Proposed Mode (Two Stream)	PSL	Full Body	94.1

5.6. Performance accuracy and state of the art comparison for the MSL dataset

The proposed model was also tested on the Mexican Sign Language (MSL) dataset, focusing on hand gesture recognition. The model's accuracy improved significantly, reaching 99.75% in the Two-Stream configuration, surpassing the 98.55% achieved by a recent Dynamic-GCN model. This highlights the proposed model's ability to capture complex sign language gestures, enhancing communication for the hearing-impaired.

TABLE 6. Performance accuracy with MSL dataset and state of the art comparison.

Method Name	Dataset Types	Joints Key points	Performance Accuracy [%]
RNN [15]	MSL	Full Body	96.44
Dynamic-GCN [14]	MSL	Full Body	98.55
Proposed Model (Single Stream)	MSL	Full Body	99.00
Proposed Mode (Two Stream)	MSL	Full Body	99.75

5.7. Performance accuracy with ASLLVD dataset

Table 8 displays the model's accuracy on the ASLLVD dataset with approximately 2700 class labels. The proposed Two-Stream GCN model set a benchmark with 34.41% accuracy, outperforming existing models like ST-GCN and 3D GCN, which achieved 16.48% and 25.05%, respectively. This performance underscores the model's effectiveness in large-scale sign language recognition.

TABLE 7. Performance accuracy and state-of-the-art comparison for the ASLLVD dataset.

Method Name	Dataset Types	Features	Class-Label	Performance Accuracy [%]
STGCN [23]	ASLLVD (ASL)	Full Body	2700	16.48
3DGCN [28]	ASLLVD (ASL)	Full Body	2700	25.05
Proposed Model (Single Stream)	ASLLVD (ASL)	Full Body	2700	31.00
Proposed Mode (Two Stream)	ASLLVD (ASL)	Full Body	2700	**34.41**

5.8. Result

In this study, we proposed an advanced Two-Stream Graph Convolution with Attention and Residual Connection (GCAR) model, utilizing a multistage method for enhanced spatial-temporal information extraction. The model incorporates joint

skeleton and motion inputs processed through sep-TCN and GCAR modules. The integration of a Channel Attention model within the GCA module refines feature extraction, significantly improving performance. Evaluated on large-scale datasets, the GCAR model demonstrated superior accuracy and computational efficiency. The results confirm that our model outperforms existing systems, indicating its potential to advance sign language recognition and improve accessibility for the hearing-impaired community.

TABLE 8. Computational complexity of the individual dataset for the proposed model.

Dataset Name	BFlops for 1 Batch (32 trial)	Total trial	No Class	Total Batch (Trial/32)	Test Batch	Computation complexity (BFlops)	Parameters (Million)
WLASL	0.13	21089	2000	659	197	25.70	0.69
MSL	0.13	3000	20	93	27	3.51	0.69
PSL	0.13	2700	19	84	25	3.27	0.69
ASLLVD	0.13	9748	2745	304	91	11.58	0.69

6. CONCLUSION

The Graph Convolution with Attention and Residual Connection (GCAR) model was introduced, featuring a two-stream multistage architecture designed for the extraction of spatial-temporal contextual information. The first stream processes joint skeleton points using Separable Temporal Convolution Networks (sep-TCN) to optimize computational efficiency. By sequentially applying Graph Convolution Networks (GCN), a deep learning module, and a channel attention module, this stream generates a refined feature representation.

Simultaneously, the second stream analyzes joint motion across consecutive frames, following a similar methodology to produce its final feature. The fusion of both streams' outputs enables effective capture of dynamic motions in sign language gestures. The integration of sep-TCN stabilizes computational complexity, while channel attention enhances the significance of non-connected skeleton points in specific spatial-temporal events. Each GCN module utilizes channel attention to extract both global and local features, contributing to model optimization.

The multistage integration of GCAR enriches spatial and temporal contextual information, producing a robust feature vector through the combination of graph and conventional CNN techniques. High accuracy across various large-scale sign language datasets demonstrates the model's effectiveness and superiority over existing methods. It also addresses challenges related to inter-gesture similarity and limited skeletal information by considering the entire body, face, and both hands, utilizing 67 key points.

Future work includes deploying this system as a practical sign language recognizer and enhancing its performance further through the incorporation of multiple camera-based recognition techniques, ensuring a reliable solution for sign language translation.

7. BIBLIOGRAPHY

[1] A. S. M. Miah, J. Shin, M. A. M. Hasan, and M. A. Rahim, “BenSignNet: Bengali sign language alphabet recognition using concatenated segmentation and convolutional neural network,” *Appl. Sci.*, vol. 12, no. 8, p. 3933, Apr. 2022.

[2] A. S. M. Miah, J. Shin, M. Al M. Hasan, M. A. Rahim, and Y. Okuyama, “Rotation, translation and scale invariant sign word recognition using deep learning,” *Comput. Syst. Sci. Eng.*, vol. 44, no. 3, pp. 2521–2536, 2023.

[3] J. Shin, A. S. Musa Miah, M. A. M. Hasan, K. Hirooka, K. Suzuki, H.-S. Lee, and S.-W. Jang, “Korean sign language recognition using transformer-based deep neural network,” *Appl. Sci.*, vol. 13, no. 5, p. 3029, Feb. 2023.

[4] A. S. M. Miah, M. A. M. Hasan, J. Shin, Y. Okuyama, and Y. Tomioka, “Multistage spatial attention-based neural network for hand gesture recognition,” *Computers*, vol. 12, no. 1, p. 13, Jan. 2023.

[5] R. Egawa, A. S. M. Miah, K. Hirooka, Y. Tomioka, and J. Shin, “Dynamic fall detection using graph-based spatial temporal convolution and attention network,” *Electronics*, vol. 12, no. 15, p. 3234, Jul. 2023.

[6] Y. Obi, K. S. Claudio, V. M. Budiman, S. Achmad, and A. Kurniawan, “Sign language recognition system for communicating to people with disabilities,” *Proc. Comput. Sci.*, vol. 216, pp. 13–20, Jan. 2023.

[7] M. A. Rahim, A. S. M. Miah, A. Sayeed, and J. Shin, “Hand gesture recognition based on optimal segmentation in human-computer interaction,” in *Proc. 3rd IEEE Int. Conf. Knowl. Innov. Invention (ICKII)*, Aug. 2020, pp. 163–166.

[8] A. S. M. Miah, M. A. M. Hasan, S. W. Jang, H. S. Lee, and J. Shin, “Multistream graph-based deep neural networks for skeleton-based sign language recognition,” *Preprint*, vol. 5, May 2023, Art. no. 2023050467.

[9] A. S. M. Miah, M. A. M. Hasan, S.-W. Jang, H.-S. Lee, and J. Shin, “Multistream general and graph-based deep neural networks for skeleton-based sign language recognition,” *Electronics*, vol. 12, no. 13, p. 2841, Jun. 2023.

[10] V. Manning, J. J. Murray, and A. Bloxs, “Linguistic human rights in the work of the world federation of the deaf,” in *The Handbook Linguistic Human Rights*, Hoboken, NJ, USA: Wiley, 2022, pp. 267–280.

[11] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu, “Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology,” *J. Educ. Comput. Res.*, vol. 58, no. 1, pp. 63–86, Mar. 2020.

[12] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, “Wearable sensor-based sign language recognition: A comprehensive review,” *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 82–97, 2021.

[13] S. Sharma and S. Singh, “Vision-based sign language recognition system: A comprehensive review,” in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Feb. 2020, pp. 140–144.

[14] A. S. M. Miah, Md. A. M. Hasan, and J. Shin, “Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model,” *IEEE Access*, vol. 11, pp. 4703–4716, 2023.

[15] K. Mejía-Peréz, D.-M. Córdova-Esparza, J. Terven, A.-M. Herrera-Navarro, T. García-Ramírez, and A. Ramírez-Pedraza, “Automatic recognition of Mexican sign language using a depth camera and recurrent neural networks,” *Appl. Sci.*, vol. 12, no. 11, p. 5523, May 2022.

[16] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, “Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video,” *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 430–439, Apr. 2018.

[17] D. Li, C. R. Opazo, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1448–1458.

[18] J. Shin, A. Matsuoka, M. A. M. Hasan, and A. Y. Srizon, “American sign language alphabet recognition by extracting feature from hand pose estimation,” *Sensors*, vol. 21, no. 17, p. 5856, Aug. 2021.