IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 May 2025, accepted 26 June 2025, date of publication 4 July 2025, date of current version 23 July 2025.

*Digital Object Identifier 10.1109/ACCESS.2025.3586096*

RESEARCH ARTICLE

# Deep Learning-Based Sign Language Recognition Using Efficient Multi-Feature Attention Mechanism

**ESMA YENISARI[1,2], (Graduate Student Member, IEEE), AND SIRMA YAVUZ[1], (Member, IEEE)**
[1]Department of Computer Engineering, Yıldız Technical University, 34349 İstanbul, Türkiye
[2]Department of Computer Engineering, Çanakkale Onsekiz Mart University, 17100 Çanakkale, Türkiye

Corresponding author: Esma Yenisari (esma.yenisari@std.yildiz.edu.tr)

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under Project 125E318.

**ABSTRACT** Sign language is a communication system used by Deaf and hard of hearing people and serves as a bridge between Deaf and hearing communities. Since sign language uses numerous visuomotor elements that include both visual perception (hand shapes, facial expressions) and physical movements (hand and arm movements), it represents a multimodal input source for Sign Language Recognition (SLR) systems. In this study, a novel deep learning-based architecture using EfficientNet and multi-feature attention mechanism is proposed to accurately recognize SL signs. Initially, general visual features are acquired through the EfficientNet model, leveraging the transfer learning paradigm. Subsequently, dataset-specific contextual features are extracted utilizing distinct network types; spatial dependencies are modeled via Convolutional Neural Networks (CNNs), whereas temporal dynamics are learned through Recurrent Neural Networks (RNNs). These features are adaptively weighted using an attention mechanism and focus on the most critical information for the classification task. This approach ensures that the most information-rich and useful components of both methods are emphasized, leading to a significant increase in final performance. Utilizing RGB video images, the proposed model, on the BosphorusSign22k General dataset comprising Turkish Sign Language (TSL) signs, achieved accuracies of 99.01% and 96.84% for sign classes of 50 and 174, respectively. Furthermore, the generalization ability of the model was demonstrated by its high accuracy of 99.84% in the Argentinian Sign Language dataset (LSA64) and 98.41% in the Indian Sign Language dataset (INCLUDE50). Experimental results indicated that the proposed model architecture has a competitive performance compared to existing SLR models reviewed in the literature.

**INDEX TERMS** Attention mechanism, computer vision, deep learning, sign language recognition, SLR datasets, vision-based recognition.

## I. INTRODUCTION

According to World Health Organization (WHO) data for the year 2025, more than 5% of the world's population (430 million people) is hearing impaired, and 34 million of these people are children [1]. It is predicted that by 2050, approximately 2.5 billion people will have varying degrees of hearing loss, and more than 700 million people, or one in 10 people, will be hearing impaired. It is clear that hearing loss is a widespread and increasing global health problem. It is evident how urgent and important it is to develop and disseminate methods that will help overcome communication barriers. Sign language (SL) is a primary means for hearing-impaired people to communicate with each other. However, if they want to communicate with someone who does not speak sign language, this situation becomes unsolvable. Understanding and using sign language is essential for hearing impaired people to participate in their education, business and daily life. SL is a basic means of communication for hearing impaired people, and the development of recognition systems has the potential to significantly improve the accessibility of this community. In this context, video-based

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio Piccinno.

SLR has become a research area of great interest in recent years. SLR systems analyze sign language to understand gestures and convert them into text or audio. Kumari and Anand [2] stated in their study that there are two categories of SLR: one is the isolated word or letter level, and the other is the continuous sentence level. Word-level SLR is especially critical for basic communication, emergency communication systems, access to educational materials, sign language training, and real-time word-level translation applications.

There are two different approaches to SLR systems: vision-based and sensor-based, depending on the type of data collection. Although sensor-based methods promise higher accuracy, they are likely to result in physical limitations and inconveniences for users. Vision-based methods are among the leading methods of sign recognition technology because of their non-invasive nature and comprehensive data collection potential. In this approach, cameras are used for image capture and analysis, whereas machine learning algorithms are often used to interpret the collected data [3]. The transition to deep learning in the field of visually based character recognition has brought about a groundbreaking change. CNNs overcome the limitations of traditional methods by learning rich spatial representations from visual inputs, whereas RNNs have played a crucial role in modeling the dynamic structure of sign language by processing these spatial features over time. However, LSTM networks have come to the fore because standard RNNs cannot cope with the gradient explosion problem for learning long sequential data. LSTMs are designed with special gate mechanisms (forget, input, output gates) to essentially solve the ''gradient vanishing'' problem. GRU, which has a simpler structure compared to LSTM, becomes quite efficient in resource-constrained environments as the model is trained faster and requires less computational resources. There are many successful studies that design SLR models using these methods alone or in combination, CNN+RNN [4], CNN+LSTM [5], in different variations in feature extraction.

Since SL is a combination of different gestures, shapes, hand-body-facial expressions, and postures known as manual and non-manual [6], visual recognition of sign language is a complex area of research in computer vision [7]. Deep learning has made significant progress in recent years in tasks such as SLR and sign language generation (SLG), but large amounts of labeled data are usually required to train these models from scratch. Furthermore, collecting and labeling sign language data can be difficult and costly. This is where transfer learning (TL) comes into play, which significantly increases the benefits of deep learning for SL. In our proposed Deep Learning-based Efficient Multi-Feature Attention Mechanism (DeepEMA) for SLR, both dataset-specific features (contextual features) and basic visual features derived from EfficientNet, a transfer learning model pre-trained on large-scale datasets, are used in the feature extraction phase. Classification is performed with deep networks using the strongest feature representations from both methods. In this way, the obtained features become parts that complement each other's weaknesses and emphasize their stronger aspects for the classification task.

This study makes the following important contributions and benefits to the field of video-based SLR. I) A novel two-channel feature learning architecture is proposed that, unlike standard two-stream architectures, simultaneously processes general visual features (via transfer learning) and dataset-specific contextual features (with CNN+RNN), while dynamically weighting the contribution of these two distinct feature sources through a source-based attention mechanism. II) To dynamically determine the importance of the extracted basic and contextual features, an attention mechanism that can weight these two different feature sources is integrated. In this way, the model is expected to focus on the most relevant information during the recognition process and improve performance. III) For the task of word-level SLR, the powerful EfficientNet-B7 model is used as a transfer learning method. IV) By compiling the studies in the literature, a diagram was created showing the different categories into which SLR systems are divided by researchers. V) The fact that the dataset used in the study consists only of RGB videos provides a practical approach that allows the system to perform signing in real-life scenarios, in the natural environment of the signers, without additional hardware or constraints. This is an important step for the widespread use of SLR systems and the development of user-friendly interfaces.

The rest of the article is organized as follows: Section II describes the previous studies in the field of SLR, their contributions to the literature, their current limitations, and the potential benefits of the proposed method. Section III provides detailed information about the datasets used in the study. Section IV provides detailed explanations of the methodology and architecture of the proposed model. Section V includes the metrics used to evaluate the model. Section VI presents the Results and Discussion, showing the performance of the proposed model on various datasets, providing a comparative analysis with existing studies. Finally, Section VII provides conclusion of the study and the proposed model, followed by outlining potential avenues for future research.

## II. RELATED WORK

SLR is a field of research that deals with the interpretation of sign language (SL) from visual or sensor-based data. Studies in this area use various techniques to analyze and recognize signed communication. These methods can be classified into different categories according to the type of sign to be recognized, such as letter/word/sentence recognition, the method of data acquisition, such as image/sensor-based, the method of data usage, such as RGB/depth/skeleton, the method of feature extraction strategy, such as manual/automatic, or the general modelling approach, such as traditional/machine learning/deep learning. Fig. 1 illustrates the variety of classifications presented in different sources for each processing

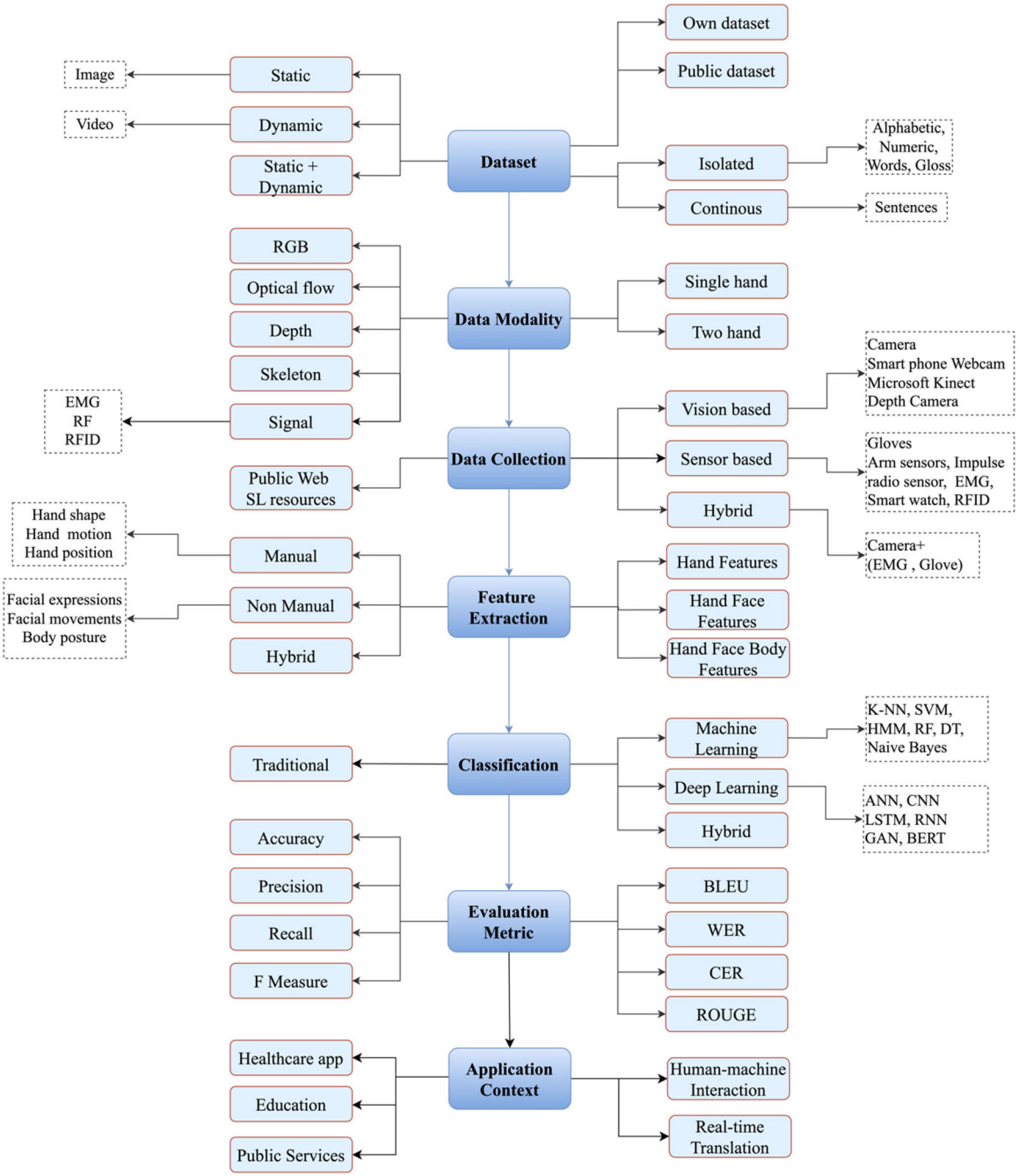


**FIGURE 1.** Taxonomy of key aspects in sign language recognition systems.

step of SLR systems, based on the literature we have reviewed.

There are two different approaches to SLR, depending on the type of data acquisition: vision-based and sensor-based.

In the vision-based approach, the input data refers to images/videos captured using methods such as camera, cell phone camera, depth camera, Kinect [8], [9], [10], [11], whereas in the sensor-based approach, the input data refers to signals obtained using methods such as smart gloves, smart watches, RFID, EMG [12], [13], [14], [15].

Yuan et al. [13] developed a hand gesture recognition system based on a wearable sensor. This study, which consists of 2 main parts, namely hardware and software, conducted data collection using the developed glove in the hardware part. The designed glove has sensors that capture fine-grained motion features such as accelerometer, bend sensor and gyroscope. Bluetooth, Wi-Fi and USB are also present in the designed glove so that the collected data can be transferred to the software. A Deep Feature Fusion Network (DFFN) model combining Deep Convolution Neural Network (DCNN) and

LSTM was proposed to combine and classify the data coming from the sensors. In the study, 5178 hand movement samples belonging to 24 English ASL letters were collected, and a precision of 99.93%was achieved with the DFFN model. 34452 hand movement samples were collected for 72 signs of CSL, and precision of 96.1% was achieved. Although sensor-based approaches are a method that is not affected by lighting conditions or cluttered backgrounds, the user must wear a device that can prevent the naturalness of hand movement [16].There are studies that use sensors on the glove to report the data generated by the hand/finger movements to the system and recognize the movement by measuring the similarity [17], [18]. However, in many cases, users refuse to wear gloves as they restrict their movement and are different from traditional interaction methods. Since all letters in TSL are gestured using two hands [19], glove sensors become even more difficult for users. Another sensor-based approach is hand movement detection studies using infrared or RFID sensors. In such studies, the movement data is captured with a sensor that is usually placed near the location where the hand movements are performed. In their study, Zhang et al. [12] proposed a model for recognizing SL based on RFID phase signals. In their study, they first calculated the environmental noise by taking the difference between the dynamic signals measured during sign language movement and the static signals measured when there was no movement. After using the Varri+ segmentation algorithm to segment the sign signals, they used the CNN and BILSTM (Bidirectional Long Short-Term Memory) model for recognition. They identified six code numbers for six different sentences belonging to daily spoken language and achieved an accuracy of 96.8% in the system trained with these six codes. Such methods require the hand to be at a certain distance from the sensor. In addition, there are many overlaps in letters and words that are similar to each other.

Vision-based systems can use cameras that are now built into many devices such as smartphones, tablets, laptops and security cameras. This makes it easier to develop and deploy SLR systems with existing infrastructure without the need for additional hardware costs and installations. It is also more functional in terms of capturing broader information. Whereas sensor-based systems typically focus on specific types of movements (e.g., hand movements, body position), image-based systems provide richer data in the form of hand shapes, hand movements, posture, facial expressions, lip movements, and contextual information (e.g., environment, other objects). Vision-based approaches require capturing images or videos of hand movements via a video camera [20]. In addition to static finger alphabets, vision-based approaches are also a good choice for dynamic recognition of words and sentences. In some of vision-based studies, hand recognition is primarily performed. Some hand recognition studies rely on segmenting images into skin and non-skin areas, using the detected skin region for identification. However, this approach often fails in the presence of dynamic backgrounds, which can interfere with accurate segmentation. Furthermore, the extensive image processing techniques frequently employed in such morphological operation-based methods can negatively impact their feasibility for real-time applications due to computational overhead.

For vision-based approaches, SLR with only RGB images is not an easy task. Therefore, researchers resort to methods based on other modalities [21]. Most studies use other image features in addition to RGB video images such as depth information [22], [23], [24], [25], [26], [27] and skeleton information [28], [29], [30]. Renjith et al. [31] conducted a study that consisted of three steps: identification of indicator frame (IIF) representing the completion of the movement, identification of key frames (IPF) representing important movements, and generation of skeleton sequence (GSS) representing the sign language word. They also used CNN to recognize and classify skeletal image sequences. They analyzed 25 samples each from the Word-Level American SL (WLASL-100), Swiss German SL (DSGS) and Indian SL (INCLUDE50) datasets and achieved 91.26%, 92.47% and 95.42%, respectively, according to the Weighted Average Accuracy (WAA) metric. In the study, they also achieved 93.26%, 93.48% and 96.64% accuracy in the respective datasets. Sincan and Keles [32] developed two models trained using motion history images (MHI) from RGB video images to perform word-level recognition. In the first model, they used RGB-MHI to focus on spatial regions, and in the second model, they used a fusion approach by combining RGB and RGB-MHI. They used the TSL datasets AUTSL and BosphorusSign22k as datasets and achieved an accuracy of 93.53% and 94.83% respectively. Khartheesvar et al. [33] achieved 87.4% and 94.8% accuracy in their study on INCLUDE dataset and its subset INCLUDE50. In their study, they extracted 258 key points from each video frame with MediaPipe holistic and used LSTM for classification. Although the use of these additional features could increase the success, each additional method requires additional computational cost. Besides additional features may not be available in many daily applications. For example, applications recorded with public internet sources, TV broadcasts or cell phone cameras are only suitable for models that can work with RGB video.

Sign language consists of manual features that express the hand and finger movements that form the basis of its movements, and non-manual features such as face, head and gestures that accompany these main features. These additional features play an important role in the recognition of SL movements [34]. All these features are present as a whole in the full frame of the video. In many studies, features such as cropped body posture, head posture, face and mouth are considered in addition to the main full frame [29], [35], [36], [37]. Although the use of additional features besides the full image in SLR has the potential to increase success, it can also bring disadvantages. The algorithms required to recognize hands in a SL video require additional processing steps and further complicate the model.

**TABLE 1.** Overview of sign language datasets.

| Dataset | Language | Level | Data Type | Class | Sample | Link |
|---|---|---|---|---|---|---|
| **Boston ASL LVD** [48] | American SL | W | Videos (different perspectives) | >3300 | 9800 | www.bu.edu/asllrp/av/dai-asllvd.html |
| **WLASL** [49] | American SL | W | Videos (public int sources) | 2000 | 21083 | https://dxli94.github.io/WLASL/ |
| **RWTH-BOSTON-50** [50] | American SL | W | Videos (different perspectives) | 50 | 483 | www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-50.php |
| **BosphorusSign22k** [51], [52] | Turkish SL | W | RGB video, depth, skeleton | 744 | 22542 | https://ogulcanozdemir.github.io/bosphorussign22k/ |
| **AUTSL** [27] | Turkish SL | W | RGB video depth, skeleton | 226 | 38336 | https://cvml.ankara.edu.tr/datasets/ |
| **INCLUDE** [53] | Indian SL | W | Videos | 263 | 4292 | https://zenodo.org/records/4010759 |
| **LSA64** [54] | Argentina SL | W | Videos | 64 | 3200 | https://facundoq.github.io/datasets/lsa64/ |
| **PSL Kinect 30** [55] | Poland SL | W | Videos & Depth From Kinect | 30 | 300 | https://vision.kia.prz.edu.pl/dynamickinect.php |
| **PSL ToF 84** [55] | Poland SL | W | Videos &Depth from ToF Cam | 84 | 1680 | https://vision.kia.prz.edu.pl/dynamictof.php |
| **DGS Kinect 40** [56] | German SL | W | Videos&Depth | 40 | 3186 | www.sign-lang.uni-hamburg.de/lrec/data/kinectdgs40.html |
| **SIGNUM** [57] | German SL | W& S | Videos | 450 W 780 S | 33210 | www.bas.uni-muenchen.de/Bas/SIGNUM/ |
| **GSL** [58] | Greek SL | W& S | RGB-Video, Depth | 310 W 331 S | 40785 10290 | https://vcl.iti.gr/dataset/gsl/ |

(W: Word Level, S: Sentence Level)

The increased computational cost and complexity cause difficulties in real-time applications and increase training and resource requirements. Since in our work the whole image is used directly without performing morphological operations on the video frames, there are significant advantages such as simplicity, low computational cost, end-to-end learning potential, data-friendliness and efficiency. Especially in resource-constrained environments, it can be assumed that only full-frame based RGB approaches will be a practical and efficient solution for rapid prototyping development processes.

Transfer learning (TL) allows models trained on large datasets (e.g. ImageNet) to capture general features These pre-trained models are effective in recognizing different objects and patterns by using its prior knowledge in the second model where there is limited data. From machine learning (ML) perspective, it can be defined as reusing the stored weights of a previously trained model to train or improve the accuracy of one's own model [38]. Das et al. [39] achieved an accuracy of 91.67% in character recognition and 97.33% in digit recognition in their study with the datasets 'Ishara-Bochon' and 'Ishara-Lipi', which consist of numbers and alphabets of Bangladeshi SL. After extracting features using the pre-trained TL architecture on the ImageNet dataset, they performed classification using the Random Forest (RF) method. In the literature, there are studies that use TL for feature extraction in SLR [38], [40], [41], studies that use deep networks [42], [43] and studies that apply deep networks after TL [44].

Recent advancements in SLR have yielded improved accuracy; however, this progress has often been accompanied by a commensurate increase in computational and storage demands. Recognizing that not all frames within SL videos contribute equally to the recognition task, Hu et al. [45]. introduced AdaSize as a strategy to mitigate computational overhead. AdaSize can model the frame resolution decision as an end-to-end learnable task. In their experiments with four large datasets, including PHOENIX14, PHOENIX14-T, CSL-Daily and CSL, they showed that AdaSize increases efficiency while reducing computational and memory requirements. Another method for reducing computational costs while increasing efficiency in SLR is the use of an attention mechanism in key frame selection [46]. Whereas conventional neural networks (NN) process the entire input equally, attention mechanisms give the network the ability to learn which parts of the input are more important for the current output. In this way, the network can produce more accurate and meaningful results by focusing on more relevant information [47].

The attention mechanism not only focuses on the important parts in the input sequences but also can be used as an intelligent feature selection mechanism by determining the relative

importance of different features extracted from a dataset [59], [60], [61]. In our proposed model, the attention mechanism is not used to directly weight the extracted features, but to determine the importance levels of the features obtained by 2 different methods. The features extracted by transfer learning and the features extracted by CNN+RNN are combined in the attention layer by weighting them according to their importance levels. This combination process allows the model to create a richer and more comprehensive feature representation. With the attention mechanism, the most useful features from both methods are weighted and the impact of unimportant features on the classification is reduced.

The most used traditional methods in dynamic hand gesture recognition systems are HMM (Hidden Markov Model), statistical methods, eigenspace-based methods, curve fitting, dynamic programming, SVM (Support Vector Machines) and active contour. Non-traditional methods used in SLR are neural network-based methods [62]. In recent years, neural networks and deep networks have been widely used in areas such as speech recognition and digital signal processing. Unlike traditional methods, deep learning-based methods do not depend on the processing of previous data and automatically extract features. Utilizing multiple processing layers, deep learning methods identify complex patterns and structures within very large datasets. This occurs through hierarchical learning, where each layer builds representations based on the concepts learned by the layer before it.

Deep learning is also invaluable in the context of big data, as it extracts high-level information from very large amounts of data [63]. Obi et al. [64] developed a desktop application that uses CNN to recognize SL letters and convert them to text. They achieved an accuracy of 96.3% for 26 letter classes from the American Sign Language (ASL) dataset. CNN is known to deliver good results for a single image. However, it is not sufficient for an input consisting of a sequence of images. Since CNN is structurally unsuitable for such a task, deep learning methods such as RNN and LSTM models, which are a type of RNN, have been developed [65].

Although it is a popular myth, there is no universal sign language [66]. Just as each country has a different spoken language, SL is also different. Although it is known that there are currently 160 different SL in the world according to data from the website Ethnologue.com [67], research to determine the SL in the world is still ongoing. SIL International [68] estimates that the actual number could exceed 400. For instance, Central Taurus Sign Language (CTSL), discovered in 2012, is a SL that occurs naturally in a remote, mountainous area in southern Turkey. It developed in three neighboring villages with little or no influence from Turkish SL [69].

Sign languages, like spoken languages, have their own grammatical structure. Most of the studies conducted in the field of SL recognition refer to the SL spoken in the country where the study was conducted or to languages in existing datasets around the world. Some of the datasets used in SLR and their summary information are listed in Table 1. The aim of this study is to develop a novel model using Turkish Sign Language (TSL) data. Whereas acknowledging the significant contributions in TSL research, we believe our proposed model can offer a complementary perspective and contribute to the ongoing exploration within the field. Recognizing the inherent value of any contribution to the TSL literature, we expect our work to be a valuable addition to this field. With this study, we aim to provide a powerful model for word-level TSL recognition, paving the way for future TSL research and applications. The model will also be validated against other country SL datasets such as LSA64 and INCLUDE50.

## III. DATASETS USED IN THIS STUDY

The proposed DeepEMA model was developed based on BosphorusSign22k and subsequently validated by testing with the LSA64 and INCLUDE50 datasets. The datasets used in the study and the summary information about these datasets are listed in Table 2.

### A. BOSPHORUSSIGN22K

22542 video recordings in the dataset, consisting of 744 signs belonging to Turkish SL, were recorded using Microsoft Kinect v2. The dataset contains RGB video, depth map and skeleton information for each sign gloss. Whereas the dataset provides RGB, depth map, and skeleton information for each sign gloss, this study focuses exclusively on RGB videos. Depth map and skeletal information were not used. The dataset is available to anyone for research purposes with the permission of the authors. It consists of 3 main categories: General, Health and Finance. There are 174 signs in the General category, 428 in the Health category, and 163 in the Finance category. The General category, which is a subset of the dataset, is also used in many scientific studies. In this study, we utilized two specific subsets derived from the BosphorusSign22k dataset for model development and evaluation, each serving a distinct purpose. The 174-class 'General' subset was selected as it is commonly used benchmark in numerous TSL research studies. Using this standard subset allows for a direct and fair comparison of our model's performance against other methods in the literature. The 50-class subset, which was randomly sampled from the 'General' category, was created to facilitate more extensive and rapid experimentation. This approach enabled us to efficiently test various architectural designs and tune hyperparameters.

### B. LSA64

The dataset contains a total of 3200 videos created by 5 repetitions for 64 signs from 10 people who have no experience with Argentine SL. The dataset, which contains both verbs and nouns, was selected from the most frequently used signs in LSA. Some of the signs are one-handed, others 2-handed.

Some recordings were made outdoors in natural light, others indoors in artificial light. In all recordings, the subjects performed the signs while standing or sitting. This variability in lighting conditions serves as a natural data

**TABLE 2.** Datasets used in this study.

| Dataset | Class | Total Sample | Signer | Video-Resolution | Frame Rate | Language |
|---|---|---|---|---|---|---|
| **BosphorusSign22k** | 744 | 22542 | 6 | 1920x1080 | 30 fps | Turkish SL |
| **BosphorusSign22k-General** | 174 | 5788 | 6 | 1920x1080 | 30 fps | Turkish SL |
| **BosphorusSign22k 50*** | 50 | 1652 | 6 | 1920x1080 | 30 fps | Turkish SL |
| **LSA64** | 64 | 3200 | 10 | 1920x1080 | 60 fps | Argentinian SL |
| **INCLUDE** | 263 | 4292 | 7 | 1920x1080 | 25 fps | Indian SL |
| **INCLUDE-50** | 50 | 958 | 7 | 1920x1080 | 25 fps | Indian SL |

*BosphorusSign22k-50 is a randomly selected subset used for initial experiments.

augmentation, testing the model's robustness. The data set is available for academic, educational or personal use. Although LSA64 encompasses fewer distinct sign classes compared to established datasets such as ASLLVD, RWTH-PHOENIX-Weather, it provides a larger number of samples per sign than many comparable SL resources.

### C. INCLUDE/INCLUDE50

The INCLUDE dataset, which consists of 4292 videos from 263 classes of Indian SL, comprises a total of 0.27 million frames. The dataset was recorded with the help of 7 high school students who were educated at a school for the deaf in SL. The videos were recorded in a bright, naturally lit classroom with no additional clothing or landscaping. The data set, selected from popular signs in ISL, includes 15 different major categories. The sign videos are 2-4 seconds long. The INCLUDE-50 dataset is a subset consisting of 50 randomly selected signs from 263 classes. They created this subset to reduce computational costs, which is a drawback when training deep learning models on a large video dataset, and to enable fast evaluations with different models. The INCLUDE-50 dataset contains 958 example videos and 60897 frames. To evaluate the proposed model, INCLUDE-50 dataset is used in this study.

## IV. METHODOLOGY

To address the potential loss of detail when resizing video frames from their original 1920 × 1080 resolution to 600 × 600, a deliberate preprocessing pipeline is used to preserve critical features. First, to mitigate the loss of important information, we did not downscale the entire frame. Instead, we began by cropping the frame to a central 1080 × 1080 area, which contains the most relevant information such as hand movements and facial expressions, while discarding irrelevant background from the edges. Second, the final 600 × 600 resolution was empirically chosen after experimenting with various input sizes (e.g., 224 × 224, 300 × 300, 456 × 456), as it yielded the optimal performance in our architecture. Most importantly, this 600 × 600 dimension was specifically selected to match the native input resolution of the pre-trained EfficientNet-B7 model. This alignment is crucial as it ensures that the powerful, pre-learned features of the model can be effectively transferred to our task. This strategy compensates for any potential loss of fine-grained detail from the resizing process and is fundamental to achieving the high recognition accuracy reported in our results

The videos in the dataset, we worked with varied in length (35-180 frames) and contained very similar frames. To standardize the variable-length videos to a fixed temporal dimension of 20 frames while preserving the most significant information, a dynamic keyframe extraction algorithm was employed. For each video, this process started with a strict similarity threshold to select only the most unique frames. The algorithm then iteratively relaxed this threshold, progressively including more frames, until exactly 20 keyframes were extracted. This adaptive approach ensures that every sample fed into the model has a uniform length, representing the most distinct moments of each sign gesture and discarding redundant information

Fig. 2 shows the architecture of the proposed DeepEMA model. For word-level sign language recognition from video, the proposed model integrates two distinct feature types: general visual features obtained via transfer learning, and contextual features extracted using a dedicated CNN+RNN component. These features go through an attention mechanism, their weights are calculated and combined according to their importance.

With this method, different features can be combined into a single representation. In this way, a richer and more robust feature space is created.

The features resulting from the attention mechanism were analyzed by different deep networks (GRU, LSTM, and BIL-

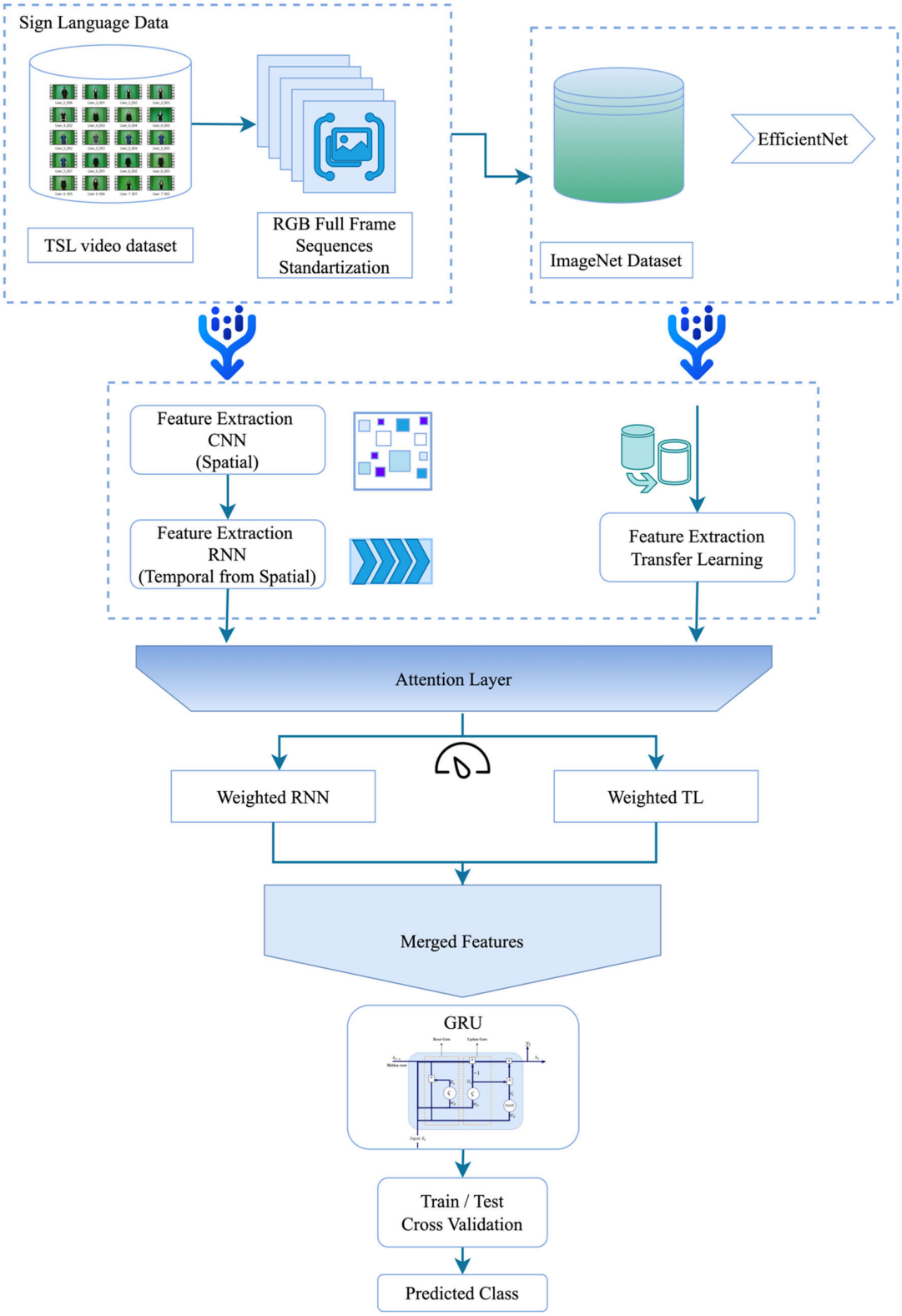


**FIGURE 2.** Architecture of proposed model.

STM) in the training phase. The best and fastest classification was obtained with GRU. The reliability of the model is increased by including the average of the results obtained in the 5-fold cross-validation in the model.

The similarity check process effectively filters out highly similar or identical images, thereby enabling the identification of genuinely unique visual content. With this approach, we captured the changes in the video content and important moments. In addition, video frames of different lengths were represented with the same number of frames.

The model extracts features from the input data in two different ways. The first are the features extracted from the model trained on ImageNet using the transfer learning method, and the second are the features we trained and extracted on our own dataset using CNN+RNN. Transfer learning allows models trained on large datasets (e.g. ImageNet) to capture general features. This pre-trained model is effective in recognizing different objects and patterns by using its prior knowledge in the second model with limited data. To improve the performance in our SLR system with limited dataset, feature extraction was performed with EfficientNet B0-B7 models trained on the ImageNet dataset. The best results were obtained with the features we extracted using EfficientNet-B7.

EfficientNet is a pure convolutional model that proposes a new scaling method that scales all depth/width/resolution dimensions equally. Tan and Le [70] systematically investigated the scaling of the model in their study. They determined that precisely optimizing network parameters such as depth, width, and resolution is key to achieving better performance. Based on this observation, they proposed a new scaling method, EfficientNet, which scales all depth/width/resolution dimensions equally, using a simple and highly effective composite scaling

*Definition 1:* If you use $2^N$ times computational resources then increase network depth by $\alpha^N$, width by $\beta^N$, image size by $\gamma^N$, where $\alpha$, $\beta$, $\gamma$ are constant coefficients on the original small model.

According to the logic of EfficientNet, the larger the input image, the more layers and channels the network should have. EfficientNet can adapt to different tasks and hardware requirements by scaling the basic dimensions of the model such as depth, width, and resolution in a balanced way. It ensures efficiency by aiming to optimize the computational cost and size of the model while increasing the accuracy of the model, which is an important feature for real-time applications such as gesture recognition. EfficientNet-B7 achieves a top 1 accuracy of 84.3% in ImageNet while being 8.4x smaller and 6.1x faster in inference than the best existing ConvNet [70]. With all these features and the ability to achieve high accuracy even with smaller datasets, EfficientNet is the TL method we used in this study.

Both spatial (with CNN) and temporal (with RNN) features are extracted from the sign language videos. Whereas CNN extracts spatial features (hand movements, facial expressions, etc.) from video images, RNN learns the sequential structure of movements by processing these features temporally. The combination of CNN+RNN, which focuses specifically on the task of action recognition, learns temporal and spatial relationships in the video. This helps to capture the dynamics and sequential structure of movements. By combining these two approaches, the model can learn both general and specific features, resulting in a richer feature representation. Deep networks such as GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory) can perform better in classifying movements by analyzing these combined features. The system learns both general basic visual and specific contextual features, leading to a better understanding of the movements.

The features obtained with transfer learning and the features obtained with CNN+RNN are combined in the attention layer by weighting them according to their importance. This combination process enables the model to create a richer and more comprehensive feature representation.

The core of this combination process is a multi-feature attention mechanism. The concept of self-attention was famously introduced in its modern form by Vaswani et al. [71] becoming the foundational component of the highly influential Transformer architecture. Inspired by this powerful principle, our DeepEMA model employs an attention mechanism specifically tailored to our architectural challenge: effectively fusing two heterogeneous feature sources.

The attention mechanism was applied separately to TL and RNN features in our study, and separate attention weights were computed for each feature type. These weights are then used to determine the importance of the relevant features. Specifically, this weighting is accomplished in a two-stage process. First, intra-sequence attention scores ($\alpha$) are computed to identify the most salient time steps within each individual feature stream. Second, a learnable, scalar source weight ($w$) is applied to each entire stream to determine its overall contribution to the final prediction. This two-level approach allows the model to learn the information from both feature types in an integrated way and achieve better performance.

The following equations describe this two-stage attention process for both the contextual ($crnn$) and transfer learning ($tl$) streams. First, the intra-sequence attention scores are computed.

Equations (1) and (2) are for the contextual (CNN+RNN) stream:

$$e_{crnn} = tanh(X_{crnn} W_{1,crnn} + b_{1,crnn}) \tag{1}$$

$$\alpha_{crnn} = softmax(e_{crnn} W_{2,crnn} + b_{2,crnn}) \tag{2}$$

Equations (3) and (4) are for the general (Transfer Learning) stream:

$$e_{tl} = tanh\left(X_{tl} W_{1,tl} + b_{1,tl}\right) \tag{3}$$

$$\alpha_{tl} = softmax\left(e_{tl} W_{2,tl} + b_{2,tl}\right) \tag{4}$$

where $X$ is the input feature sequence, $W_1$, $W_2$ are learnable weight matrices, and $b_1$, $b_2$ are learnable bias vectors.

In the second stage, the model learns the overall reliability of each feature source through a single, learnable scalar parameter, ($w$). These weights, $w_{crnn}$ and $w_{tl}$ are optimized during training via backpropagation, allowing the model to

decide which stream's information is more valuable for the classification task.

The final output of the attention layer is produced by weighting the original features with both the intra-sequence attention scores ($\alpha$) and the source-specific reliability weight ($w$). This process yields the "Weighted RNN" and "Weighted TL" outputs shown in (5) (6).

$$CR_{weighted} = X_{crnn} \odot \alpha_{crnn} \odot w_{crnn} \tag{5}$$

$$TL_{weighted} = X_{tl} \odot \alpha_{tl} \odot w_{tl} \tag{6}$$

Here, $\odot$ denotes the element-wise multiplication. These two weighted feature streams are then concatenated to form a single, enriched feature vector that is passed to the classification head.

Then weighted features are combined. The combined features are used as input for the deep model (LSTM, GRU, BILSTM). This model classifies SL gestures by analyzing the features. The model was trained and evaluated using 5-fold cross-validation. With cross-validation, it is possible to evaluate how the model performs on different data segments. Finally, the model outputs a prediction indicating the specific SL class to which the input video corresponds.

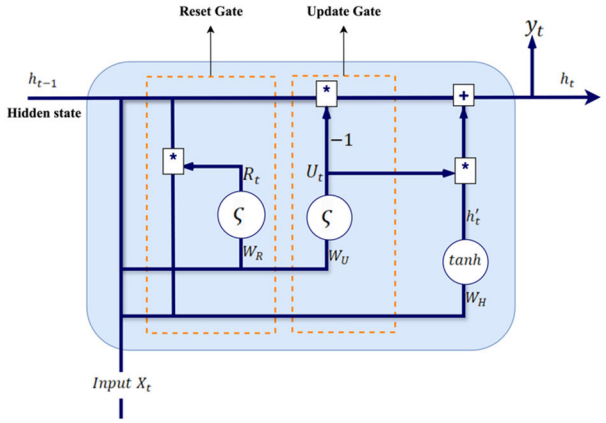


**FIGURE 3.** Gated recurrent unit (GRU) block.

The best results were obtained with the GRU-based classifier from deep networks fed with the obtained dual features. GRU was first introduced by Cho et al. [72] and is a type of RNN architecture that addresses the problem of short-term memory. GRU is an RNN method used in modeling sequential data when computational resources are limited. GRU offers a simpler structure than LSTM. The GRU architecture simplifies the gating mechanism by merging the separate input and forget gates of an LSTM into a single update gate. GRU, which operates with a simpler structure and without the high computational cost of LSTM without sacrificing performance, provides successful results for many sequential tasks. As shown in Fig. 3, GRU consists of two main gates (Update, Reset). The reset gate determines how much of the information in the previous hidden layer is forgotten (7). The update gate determines how much of the input data is used to update the hidden layer (8).

$$R_t = sigmoid\,(W_r \cdot [h_{t-1}, x_t]) \tag{7}$$

$$U_t = sigmoid\,(W_u \cdot [h_{t-1}, x_t]) \tag{8}$$

$x_t$ is the current input in the form of a sequence with which the GRU is fed, $h_{t-1}$ is the hidden state in which the recurrent calculation is performed, $W$ is the weight matrix learned during training.

$h'_t$ is the hidden candidate state (9) and $h_t$ is the new hidden state (10) computed in the update gate by weighing the previous hidden state and the hidden candidate state.

$$h'_t = tanh(W_h \cdot [R_t \cdot h_{t-1}, x_t]) \tag{9}$$

$$h_t = (1 - U_t) \cdot h_{t-1} + U_t \cdot h'_t \tag{10}$$

$y_t$ the output of the GRU, which is generated in the output layer depending on the desired task (single number, sequence or probability distribution).

GRU uses special gates that adaptively control the updating and storage of information within the unit. By managing this information flow, the gates ensure that the gradients remain informative during backpropagation over long sequences. This allows the network to selectively remember important past contexts while incorporating new relevant inputs.

## V. EVALUATION METRICS

There are many criteria that can be used to evaluate the performance of SLR model. For isolated SL, the commonly used evaluation metric is Accuracy rates [73]. Precision, Recall and F-score are other commonly used metrics besides accuracy. For the proposed system, accuracy is the chosen criterion to make comparisons with other studies in the literature. The performance metrics we used in this study are given in (11), (12), (13), and (14).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$FScore = \frac{2TP}{2TP + FP + FN} \tag{14}$$

where $TP$ is True Positive, $TN$ is True Negative, $FP$ is False Positive and $FN$ is False Negative

## VI. EXPERIMENTAL RESULTS AND DISCUSSION

Utilizing only RGB information from sign language videos, we first performed frame elimination based on a similarity threshold. Subsequently, we extracted dual features—both specialized (via a CNN+RNN architecture) and generalized (via Transfer Learning)—from this filtered data. These dual features were then weighted using a Self-Attention mechanism before being fed into a GRU deep neural network. Thus, before reaching the classification layer with GRU, we extracted more meaningful and summarizing features from the data we had at each stage.

**TABLE 3.** Performance results of the proposed model on the BosphorusSign dataset.

| Class | Method | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|---|
| 50 | TL (EfficientNet) GRU | 97.58 | 97.96 | 97.53 | 97.51 |
| 50 | CNN+RNN GRU | 96.31 | 96.23 | 96.25 | 95.78 |
| 50 | Multi-Features GRU | 97.94 | 98.25 | 97.99 | 97.93 |
| 50 | **Multi-Features Attention GRU** | **99.01** | 99.34 | 99.15 | 99.19 |
| 174 | Multi-Features GRU | 95.92 | 95.52 | 95.86 | 95.42 |
| 174 | **Multi-Features Attention GRU** | **96.84** | 97.30 | 96.58 | 96.37 |

The proposed model was evaluated in 3 different sign languages and comparison tables of previous studies conducted on the same datasets are presented together. Although the datasets examined are the same, the parts of the dataset used as input in the studies (skeletal joints, hand posture, heat maps) may differ. In the studies, the BosphorusSign22k dataset was first used to develop the model. The model was then evaluated on the LSA64 and Include50 datasets.

The studies carried out with BosphorusSign22k and the values resulting from the 5-fold average values are listed in Table 3. The table provides a comparative evaluation of the hand gesture recognition models that we tested to investigate the effects of different architectural choices. The comparison is performed for two different test setups defined by the number of gesture classes: 50 and 174. For each model, the table reports the test accuracy, precision, recall, and F-score to allow a comprehensive analysis of the performance characteristics.

On the 50-class dataset, we first evaluated the performance using individual feature streams. The model utilizing TL features achieved a respectable test accuracy of 97.58%. The model employing features extracted via the CNN+RNN architecture yielded a test accuracy of 96.31%. Combining both feature types 'Multi-Features' demonstrated a synergistic effect, improving the test accuracy to 97.94%. Notably, the most significant performance gain was observed when incorporating an attention mechanism with the multi-feature approach. Proposed DeepEMA model, 'Multi-Features + Attention + GRU' configuration achieved the highest performance on this dataset, reaching a test accuracy of 99.01%. This highlights the efficacy of the attention mechanism in selectively focusing on the most informative parts of the combined feature representations. The experiments were extended to a more challenging scenario using the 174-class dataset. The proposed model obtained a test accuracy of 96.84%.

Table 4 presents a comparative analysis of the proposed DeepEMA model's performance against several state-of-the-art methods conducted on the BosphorusSign22k dataset. The table details the publication year, reference, input modality, number of classes, technique employed, and the reported accuracy (%) for each study. As observed, the input modalities vary significantly across the compared methods, ranging from heat maps and hand shape to more complex combinations involving RGB data, pose information (body, finger, face), optical flow, hand skeleton joints, and facial expressions combined with hand gestures. The proposed DeepEMA model, evaluated on RGB full frames, demonstrates notable accuracy compared to the existing literature. Specifically, when trained and tested on a dataset with 50 classes and 1652 samples, our model achieves an accuracy of 99.01%. Even when challenged with a larger and more complex dataset comprising 174 classes and 5788 samples, DeepEMA maintains a competitive accuracy of 96.84%, outperforming the other listed methods. The high accuracy achieved across different dataset scales also suggests the robustness and generalizability of our approach.

Table 5 compares the proposed model with other studies conducted with the LSA64 dataset, their input modality and the techniques used. It shows that the proposed model is more successful than the other compared methods with 99.84% accuracy.

LSA64 dataset contains videos recorded under varied lighting conditions (both indoor artificial and outdoor natural light). This result highlights the model's inherent robustness and its ability to generalize across challenging real-world scenarios without explicit light normalization preprocessing. Furthermore, this high level of success is likely supported by the LSA64 dataset's structure, which offers a relatively large number of training samples per class (3200 samples for 64 signs) compared to some other datasets, potentially enabling more robust learning. This finding serves as further evidence that a greater number of examples per sign facilitates more effective model training and generalization in sign language recognition.

**TABLE 4.** Accuracy comparison with advanced methods for BosphorusSign dataset.

| Year | Reference | Input Modality | Class | Method | Acc % |
|---|---|---|---|---|---|
| **2019** | [74] | Heat maps skeleton joints from sensor. Hand shape | 174 | Temporal Accumulative Features | 81.58 |
| **2021** | [75] | RGB, Pose, optical flow | 174 | Inception 3D LSTM-RNN | 89.35 |
| **2024** | [76] | Finger, full-body, face features | 174 | MultiChannel MobileNetV2 | **97.15** |
| **2024** | [77] | Hand skeleton Hand Joint connections | 744 | Graph Conv. Networks (GCNs) | 89.67 |
| **2024** | [78] | Facial expressions Hand gestures | 744 | ResNet-18 LSTM MSE | 94 |
| **2025** | Proposed DeepEMA | RGB full frame | 50 | EfficientNet MultiFeature Attention GRU | 99.01 |
| **2025** | Proposed DeepEMA | RGB full frame | 174 | EfficientNet MultiFeature Attention GRU | 96.84 |

**TABLE 5.** Accuracy Comparison with advanced methods for LSA64 dataset.

| Year | Reference | Input Modality | Class | Method | Acc % |
|---|---|---|---|---|---|
| **2016** | [79] | RGB-Hand | 64 | BoW + SubCls | 91.70 |
| **2022** | [80] | RGB Full Frame | 64 | Pruned VGG | 95.50 |
| **2022** | [80] | RGB Full Frame | 64 | InceptionV3-GRU | 74.22 |
| **2024** | [81] | Hand Face Keypoints | 64 | Transformer | 98.25 |
| **2024** | [76] | Finger Body Face Pose | 64 | MultiChannel-MobileNetV2, | 99.78 |
| **2025** | Proposed DeepEMA | RGB Full Frame | 64 | EfficientNet MultiFeature Attention GRU | **99.84** |

Table 6 presents a comparative analysis of hand gesture recognition accuracy achieved by various state-of-the-art methods and the proposed DeepEMA model. The table details the publication year, reference, input modality, number of classes, sample size, technique employed, and the reported accuracy (%) for each study. All compared methods, except [86], were evaluated on a dataset with 50 gesture classes and 958 samples, allowing for a direct comparison under similar experimental conditions. The surveyed literature demonstrates a variety of input modalities and techniques employed for hand gesture recognition. Methods utilizing pose information, either as video or hand pose, achieve competitive accuracies, as seen in [53] with 94.5% in [87] with 97.44%. RGB video-based approaches also show promising results, with [82] reporting 91.30% and [84] achieving 92.46% using advanced temporal modeling with attention mechanisms. Notably, [86], although using a significantly larger and more complex dataset, reported an accuracy of 96% using transfer learning with a deep learning CNN. The proposed DeepEMA model achieves a state-of-the-art accuracy of 98.41% on the 50-class dataset with 958 samples.

The integration of multi-feature extraction and the attention mechanism within the GRU network appears to be highly effective in capturing the relevant spatio-temporal features for accurate hand gesture recognition.

Our work builds upon the established multi-stream paradigm demonstrated in studies like [75] and [76]. However, its primary novelty lies in the specific fusion strategy. We introduce a two-stage attention mechanism where attention is first applied within each feature stream to identify salient temporal information. Subsequently, a learnable scalar weight is applied to each entire stream to adaptively control the overall contribution of each feature channel to the final prediction.

To evaluate the effectiveness of our proposed DeepEMA model in a broader context, we compare its performance against several state-of-the-art methods reported in the literature on various sign language recognition datasets. Table 7 summarizes this comparison, detailing the dataset characteristics (number of classes, samples), input modalities, techniques employed, and reported accuracies alongside the results obtained by our model on the BosphorusSign22k, LSA64, and Include50 datasets.

**TABLE 6.** Accuracy comparison with advanced methods for INCLUDE dataset.

| Year | Reference | Input Modality | Class | Method | Acc % |
|---|---|---|---|---|---|
| **2020** | [53] | Pose video | 50 | MobileNetV2 BILSTM | 94.5 |
| **2023** | [82] | RGB videos | 50 | DCDW-LSTM | 91.30 |
| **2023** | [83] | Mediapipe Pose Hands | 50 | BERT Transformer | 89.5 |
| **2024** | [84] | RGB videos | 50 | C3D-BiLSTM MHAttention | 92.46 |
| **2023** | [85] | RGB videos Full frame | 50 | Meta-Learning + LSTM RNN | 84.52 |
| **2023** | [86] | RGB videos | 263 | Transfer Learning Deep Learning | 96 |
| **2024** | [87] | Hand Pose | 50 | MediaPipe CNN BILSTM | 97.44 |
| **2024** | [88] | Mediapipe Pose Face Hands | 50 | Mediapipe Attention-based BILSTM | 88.63 |
| **2024** | [33] | Pose, Hand, Body | 50 | MediaPipe LSTM | 94.8 |
| **2025** | Proposed DeepEMA | RGB video frames | 50 | EfficientNet MultiFeature Attention GRU | **98.41** |

**TABLE 7.** Accuracy comparison with advanced methods for different datasets with similar scale.

| Dataset | Reference | Input Modalitiy | Class | Sample | Method | Acc % |
|---|---|---|---|---|---|---|
| **ISL Words** | [10] | RGB videos | 17 | 340 | Hybrid LSTM | 83.36 |
| **IISL2020** | [42] | RGB videos | 11 | 1160 | LSTM-GRU | 97.1 |
| **GSL** | [89] | 3D skeletal hand and multi-body poses. | 20 | ~840 | End-to-end Fourier CNN (EFCNN). | 90.69 |
| **DArSL50** | [43] | hand and body key points | 50 | 7500 | LSTM with an attention mech. | 85 |
| **BSL-38** | [8] | RGB images | 38 | 12160 | Graph meets Attention and CNN (GmTC) | 93 |
| **KSU-SSL** | [9] | RGB videos | 40 | 16000 | CNN-BILSTM | 94.46 |
| **FSL** | [22] | Hand movements, head pose, gaze direction | 33 | 1188 | HMM | 93.87 |
| **Bosphorus Sign22k** | Proposed DeepEMA | RGB videos | 50 | 1652 | EfficientNet MultiFeature Attention GRU | 99.01 |
| **LSA64** | Proposed DeepEMA | RGB videos | 64 | 3200 | EfficientNet MultiFeature Attention GRU | 99.84 |
| **Include50** | Proposed DeepEMA | RGB videos | 50 | 958 | EfficientNet MultiFeature Attention GRU | 98.41 |

## VII. CONCLUSION

In this study, a novel deep learning approach for video-based sign language recognition is presented. The proposed model uses both general (obtained by transfer learning) and dataset-specific (extracted by CNN+RNN) feature representations via a two-channel architecture to recognize signs at the word level. In the existing literature, the sequential use of feature extraction methods is common. In our model, the aim is to increase the performance by using feature extractors simultaneously rather than sequentially. A key aspect of the proposed model is the role of its attention mechanism, which dynamically weights the contributions of the general and specific feature streams. This approach to source-specific weighting for multi-feature fusion is a central contribution of our work. Furthermore, this study investigated the application of the EfficientNet-B7 model for word-level sign language recognition, evaluating its effectiveness in this context.

The integration of general features obtained by transfer learning with spatial and temporal features extracted from CNN and RNN layers demonstrated the potential of the

model to create a more powerful, robust and effective model for different types of data. This integrated approach aims to overcome the limitations of the features used alone and eliminate the mutual shortcomings. The analyses performed indicated that promising results can be achieved in this direction. The proposed model was tested on datasets in different languages with different sized classes and its performance was verified. The model obtained 96.84% accuracy, 97.30% precision, 96.58% recall, and 96.37% F-score on the BosphorusSign22k General dataset of TSL. In addition, the model achieved accuracies of 99.84%, and 98.41% on the LSA-64 and INCLUDE50 datasets, respectively, which were the highest values found in the literature we reviewed. Since the datasets used in the study consist only of RGB videos, the model, when developed and used in real-life scenarios, would allow signers to perform their signing in their natural environment without additional hardware or constraints. This increases the practical applicability and scalability of the proposed system.

In future studies, the performance of the proposed model on different and more extensive sign language datasets will be investigated in detail. In addition, we aim to further increase the recognition accuracy of the model on larger datasets by investigating different variants of the attention mechanism and merging strategies. A crucial direction for future work will be a direct benchmark of the DeepEMA model against Transformer-based architectures. Such a study would provide valuable insights into the performance and efficiency trade-offs between our hybrid CNN+RNN approach and the self-attention mechanisms central to Transformers. Furthermore, real-time applications of the model and its performance in resource-constrained environments will be among the future research topics.

## REFERENCES

[1] World Health Organization (WHO). *Deafness and Hearing Loss*. Accessed: Mar. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] D. Kumari and R. S. Anand, "Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism," *Electronics*, vol. 13, no. 7, p. 1229, Mar. 2024.

[3] A. Osman Hashi, S. Zaiton Mohd Hashim, and A. Bte Asamah, "A systematic review of hand gesture recognition: An update from 2018 to 2024," *IEEE Access*, vol. 12, pp. 143599–143626, 2024, doi: 10.1109/ACCESS.2024.3421992.

[4] R. C. Poonia, "LiST: A lightweight framework for continuous Indian sign language translation," *Information*, vol. 14, no. 2, p. 79, Jan. 2023.

[5] N. Basnin, L. Nahar, and M. S. Hossain, "An integrated CNN-LSTM model for Bangla lexical sign language recognition," in *Proc. Int. Conf. Trends Comput. Cogn. Eng. (TCCE)*, Singapore, Dec. 2020, pp. 695–707.

[6] B. A. Al Abdullah, G. A. Amoudi, and H. S. Alghamdi, "Advancements in sign language recognition: A comprehensive review and future prospects," *IEEE Access*, vol. 12, pp. 128871–128895, 2024, doi: 10.1109/ACCESS.2024.3457692.

[7] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 113794.

[8] A. S. M. Miah, M. A. M. Hasan, Y. Tomioka, and J. Shin, "Hand gesture recognition for multi-culture sign language using graph and general deep learning network," *IEEE Open J. Comput. Soc.*, vol. 5, pp. 144–155, 2024, doi: 10.1109/OJCS.2024.3370971.

[9] L. Al Khuzayem, S. Shafi, S. Aljahdali, R. Alkhamesie, and O. Alzamzami, "Efhamni: A deep learning-based Saudi sign language recognition application," *Sensors*, vol. 24, no. 10, p. 3112, May 2024.

[10] A. Venugopalan and R. Reghunadhan, "Applying hybrid deep neural network for the recognition of sign language words used by the deaf COVID-19 patients," *Arabian J. Sci. Eng.*, vol. 48, no. 2, pp. 1349–1362, Feb. 2023.

[11] P. Dreuw, J. Förster, T. Deselaers, and H. Ney, "Efficient approximations to model-based joint tracking and recognition of continuous sign language," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.

[12] Y. Zhang, Y. Wang, F. Li, W. Yu, C. Wang, and Y. Jiang, "Sign language recognition based on CNN-BiLSTM using RF signals," *IEEE Access*, vol. 12, pp. 190487–190504, 2024, doi: 10.1109/ACCESS.2024.3517417.

[13] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, and L. Yuan, "Hand gesture recognition using deep feature fusion network based on wearable sensors," *IEEE Sensors J.*, vol. 21, no. 1, pp. 539–547, Jan. 2021.

[14] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, Oct. 2017.

[15] B. Demircioglu, G. Bülbül, and H. Köse, "Turkish sign language recognition with leap motion," in *Proc. 24th Signal Process. Commun. Appl. Conf. (SIU)*, May 2016, pp. 589–592.

[16] B. Baatar and J. Tanaka, "Comparing sensor based and vision based techniques for dynamic gesture recognition," in *Proc. 10th Asia Pacific Conf. Comput. Human Interact. (APCHI)*, 2012, pp. 2–21.

[17] P. Pandey and V. Jain, "Hand gesture recognition for sign language recognition: A review," *Int. J. Sci. Eng. Technol. Res. (IJSETR)*, vol. 4, no. 3, pp. 466–467, Jan. 2015.

[18] N. AdnanIbraheem and R. Zaman Khan, "Survey on various gesture recognition technologies and techniques," *Int. J. Comput. Appl.*, vol. 50, no. 7, pp. 38–44, Jul. 2012.

[19] TSL. (2024). *The Contemporary Turkish Sign Language Dictionary*. Accessed: Apr. 2025. [Online]. Available: https://tidsozluk.aile.gov.tr/en/

[20] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131–153, Aug. 2019.

[21] N. Sarhan and S. Frintrop, "Transfer learning for videos: From action recognition to sign language recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1811–1815.

[22] M. Jebali, A. Dakhli, and M. Jemni, "Vision-based continuous sign language recognition using multimodal sensor fusion," *Evolving Syst.*, vol. 12, no. 4, pp. 1031–1044, Dec. 2021.

[23] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3408–3418.

[24] A. A. I. Sidig, H. Luqman, S. Mahmoud, and M. Mohandes, "KArSL: Arabic sign language database," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, Jan. 2021.

[25] P. Dreuw, P. Steingrube, T. Deselaers, and H. Ney, "Smoothed disparity maps for continuous American sign language recognition," in *Proc. 4th Iberian Conf. Pattern Recognit. Image Anal. (IbPRIA)*, Póvoa de Varzim, Portugal, Jan. 2009, pp. 24–31.

[26] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1114–1119.

[27] O. M. Sincan and H. Y. Keles, "AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181340–181355, 2020.

[28] R. Li and L. Meng, "Multi-view spatial–temporal network for continuous sign language recognition," 2022, *arXiv:2204.08747*.

[29] Z. Zhou, V. W. L. Tam, and E. Y. Lam, "A cross-attention BERT-based framework for continuous sign language recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 1818–1822, 2022.

[30] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 17043–17056.

[31] S. Renjith, M. S. Sumi Suresh, and M. Rashmi, "An effective skeleton-based approach for multilingual sign language recognition," *Eng. Appl. Artif. Intell.*, vol. 143, Mar. 2025, Art. no. 109995.

[32] O. Mercanoglu Sincan and H. Y. Keles, "Using motion history images with 3D convolutional networks in isolated sign language recognition," *IEEE Access*, vol. 10, pp. 18608–18618, 2022.

[33] G. Khartheesvar, M. Kumar, A. K. Yadav, and D. Yadav, "Automatic Indian sign language recognition using MediaPipe holistic and LSTM network," *Multimedia Tools Appl.*, vol. 83, no. 20, pp. 58329–58348, Dec. 2023.

[34] S. Alyami, H. Luqman, and M. Hammoudeh, "Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects," *Inf. Process. Manage.*, vol. 61, no. 5, Sep. 2024, Art. no. 103774.

[35] W. Aditya, T. K. Shih, T. Thaipisutikul, A. S. Fitriajie, M. Gochoo, F. Utaminingrum, and C.-Y. Lin, "Novel spatio-temporal continuous sign language recognition using an attentive multi-feature network," *Sensors*, vol. 22, no. 17, p. 6452, Aug. 2022.

[36] P. Jiao, Y. Min, Y. Li, X. Wang, L. Lei, and X. Chen, "CoSign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20619–20629.

[37] F. Wei and Y. Chen, "Improving continuous sign language recognition with cross-lingual signs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23555–23564.

[38] M. Zakariah, Y. A. Alotaibi, D. Koundal, Y. Guo, and M. Mamun Elahi, "Sign language recognition for Arabic alphabets using transfer learning technique," *Comput. Intell. Neurosci.*, vol. 2022, Apr. 2022, Art. no. 4567989.

[39] S. Das, M. S. Imtiaz, N. H. Neom, N. Siddique, and H. Wang, "A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118914.

[40] M. K. Vidhyalakshmi, K. Jagadeesh, B. S. Dharmaraj, P. Sankar N, and R. G. Kumar, "Indian sign language recognition using transfer learning with efficient net," *Smart Sci.*, vol. 12, no. 2, pp. 269–280, Apr. 2024.

[41] B. Mocialov, G. Turner, and H. Hastie, "Transfer learning for British sign language modelling," in *Proc. 5th Workshop NLP Similar Lang.*, Jan. 2020, pp. 101–110.

[42] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.-B. Gil-González, and J. M. Corchado, "Deepsign: Sign language detection and recognition using deep learning," *Electronics*, vol. 11, no. 11, p. 1780, Jun. 2022.

[43] R. S. Abdul Ameer, M. A. Ahmed, Z. T. Al-Qaysi, M. M. Salih, and M. L. Shuwandy, "Empowering communication: A deep learning framework for Arabic sign language recognition with an attention mechanism," *Computers*, vol. 13, no. 6, p. 153, Jun. 2024.

[44] Q. Tian, W. Sun, L. Zhang, H. Pan, Q. Chen, and J. Wu, "Gesture image recognition method based on DC-Res2Net and a feature fusion attention module," *J. Vis. Commun. Image Represent.*, vol. 95, Sep. 2023, Art. no. 103891.

[45] L. Hu, L. Gao, Z. Liu, and W. Feng, "Scalable frame resolution for efficient continuous sign language recognition," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109903.

[46] W. Pan, X. Zhang, and Z. Ye, "Attention-based sign language recognition network utilizing keyframe sampling and skeletal features," *IEEE Access*, vol. 8, pp. 215592–215602, 2020.

[47] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, Dec. 2022.

[48] C. Neidle, A. Thangali, and S. Sclaroff, "Challenges in development of the American sign language lexicon video dataset (ASLLVD) corpus," in *Proc. 5th Workshop Represent. Process. Sign Lang., Interact. Corpus Lexicon, Lang. Resour. Eval. Conf. (LREC)*, 2012, pp. 1–6.

[49] 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.

[50] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "Combination of tangent distance and an image distortion model for appearance-based sign language recognition," in *Pattern Recognition* (Lecture Notes in Computer Science), vol. 3663. Berlin, Germany: Springer, 2005, pp. 401–408.

[51] O. Özdemir, A. A. Kındıroğlu, N. C. Camgöz, and L. Akarun, "BosphorusSign22k sign language recognition dataset," in *Proc. 9th Workshop Represent. Process. Sign Lang. (LREC)*, Jan. 2020, pp. 181–188.

[52] N. C. Camgöz, A. A. Kındıroğlu, S. Karabüklü, M. Kelepir, A. S. Özsoy, and L. Akarun, "BosphorusSign: A Turkish sign language recognition corpus in health and finance domains," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, May 2016, pp. 1383–1388.

[53] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "INCLUDE: A large scale dataset for Indian sign language recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1366–1375.

[54] F. F. Ronchetti, F. Quiroga, C. A. Estrebou, L. C. Lanzarini, and A. Rosete, "LSA64: An Argentinian sign language dataset," in *Proc. 22nd Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016, p. 5.

[55] T. Kapuscinski, M. Oszust, M. Wysocki, and D. Warchol, "Recognition of hand gestures observed by depth cameras," *Int. J. Adv. Robotic Syst.*, vol. 12, no. 4, p. 36, Apr. 2015.

[56] H. Cooper, E. J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2205–2231, 2012.

[57] U. V. Agris and K. Kraiss, "SIGNUM database: Video corpus for signer-independent continuous sign language recognition," in *Proc. 4th Workshop Represent. Process. Sign Lang.: Corpora Sign Lang. Technol.*, Jan. 2010, pp. 243–246.

[58] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. Th. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 1750–1762, 2022.

[59] S. Das, S. K. Biswas, and B. Purkayastha, "An expert system for Indian sign language recognition using spatial attention–based feature and temporal feature," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 3, pp. 1–23, Mar. 2024.

[60] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2822–2832, Sep. 2019.

[61] J. Zhang, Q. Wang, Q. Wang, and Z. Zheng, "Multimodal fusion framework based on statistical attention and contrastive attention for sign language recognition," *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1431–1443, Feb. 2023.

[62] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Comput. Vis. Image Understand.*, vol. 141, pp. 152–165, Dec. 2015.

[63] N. Rusk, "Deep learning," *Nat. Methods*, vol. 13, no. 1, p. 35, Jan. 2016.

[64] 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, May 2023.

[65] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[66] K. Emmorey, "Ten things you should know about sign languages," *Current Directions Psychol. Sci.*, vol. 32, no. 5, pp. 387–394, Oct. 2023.

[67] *Ethnologue*. Accessed: Mar. 2025. [Online]. Available: https://www.ethnologue.com/subgroup/2/

[68] SIL International. (2025). *Sign Languages*. Accessed: Mar. 2025. [Online]. Available: https://www.sil.org/sign-languages

[69] R. Ergin and D. Brentari, "Handshape preferences for objects and predicates in central Taurus sign language," in *Proc. 41st Annu. Boston Univ. Conf. Lang. Develop. (BUCLD)*, Somerville, MA, USA: Cascadilla Press, 2017, pp. 222–235.

[70] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jan. 2019, pp. 6105–6114.

[71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, Jun. 2017.

[72] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. SSST-8, 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014, pp. 103–111.

[73] T. Tao, Y. Zhao, T. Liu, and J. Zhu, "Sign language recognition: A comprehensive review of traditional and deep learning approaches, datasets, and challenges," *IEEE Access*, vol. 12, pp. 75034–75060, 2024, doi: 10.1109/ACCESS.2024.3398806.

[74] A. A. Kindiroglu, O. Özdemir, and L. Akarun, "Temporal accumulative features for sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1288–1297.

[75] C. GüNdüZ and H. Polat, "Turkish sign language recognition based on multistream data fusion," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 29, no. 2, pp. 1171–1186, Mar. 2021.
[76] A. Akdag and O. K. Baykan, "Multi-stream isolated sign language recognition based on finger features derived from pose data," *Electronics*, vol. 13, no. 8, p. 1591, Apr. 2024.
[77] O. Özdemir, İ. M. Baytaş, and L. Akarun, "Hand graph topology selection for skeleton-based sign language recognition," in *Proc. IEEE 18th Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2024, pp. 1–5.
[78] İ. M. Baytaş and I. Erdoğan, "Signer-independent sign language recognition with feature disentanglement," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 32, no. 3, pp. 420–435, May 2024.
[79] F. Ronchetti, F. Quiroga, C. A. Estrebou, L. C. Lanzarini, and A. R. Suárez, "Sign languague recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language," in *Proc. Ibero-Am. Conf. Artif. Intell. (IBERAMIA)*. Cham, Switzerland: Springer, Jan. 2016, pp. 338–349.
[80] M. Marais, D. Brown, J. Connan, A. Boby, and L. L. Kuhlane, "Investigating signer-independent sign language recognition on the LSA64 dataset," in *Proc. South. Afr. Telecommun. Netw. Appl. Conf. (SATNAC)*, Sep. 2022, pp. 1–11.
[81] S. Alyami, H. Luqman, and M. Hammoudeh, "Isolated Arabic sign language recognition using a transformer-based model and landmark keypoints," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 1, pp. 1–19, Jan. 2024.
[82] A. Dey and S. Biswas, "Gesture recognition for ISL-question signs in videos using DCDW-LSTM attention," in *Proc. IEEE 5th PhD Colloq. Emerg. Domain Innov. Technol. Soc. (PhD EDITS)*, Nov. 2023, pp. 1–2.
[83] K. B. Prathap, G. D. Swaroop, B. P. Kumar, V. Kamble, and M. Parate, "ISLR: Indian sign language recognition," in *Proc. 2nd Int. Conf. Paradigm Shifts Commun. Embedded Syst., Mach. Learn. Signal Process. (PCEMS)*, Apr. 2023, pp. 1–6.
[84] A. Dey, S. Biswas, and D.-N. Le, "Recognition of wh-question sign gestures in video streams using an attention driven C3D-BiLSTM network," *Proc. Comput. Sci.*, vol. 235, pp. 2920–2931, Jun. 2024.
[85] R. K. Katti and P. Desai, "Character and word level gesture recognition of Indian sign language," in *Proc. IEEE 8th Int. Conf. Converg. Technol. (I2CT)*, Apr. 2023, pp. 1–6.
[86] M. Mahyoub, F. Natalia, S. Sudirman, and J. Mustafina, "Sign language recognition using deep learning," in *Proc. Int. Conf. Develop. eSyst. Eng. (DeSE)*, Jan. 2023, pp. 184–189.
[87] H. M. A. Ghanimi, S. Sengan, V. B. Sadu, P. Kaur, M. Kaushik, R. Alroobaea, A. M. Baqasah, M. Alsafyani, and P. Dadheech, "An open-source MP+ CNN+ BiLSTM model-based hybrid model for recognizing sign language on smartphones," *Int. J. Syst. Assurance Eng. Manage.*, vol. 15, no. 8, pp. 3794–3806, Aug. 2024.
[88] C. Sujatha, P. Jadi, N. B. Shubham, S. N. Habib, U. M. Chaitanya, and P. Desail, "Improved Indian regional sign language recognition with extended IRKSL dataset," in *Proc. 6th Int. Conf. Comput. Intell. Netw. (CINE)*, Dec. 2024, pp. 1–6.
[89] S. B. Abdullahi, K. Chamnongthai, V. Bolon-Canedo, and B. Cancela, "Spatial–temporal feature-based end-to-end Fourier network for 3D sign language recognition," *Expert Syst. Appl.*, vol. 248, Aug. 2024, Art. no. 123258.

**ESMA YENISARI** (Graduate Student Member, IEEE) received the M.S. degree in computer engineering from Pamukkale University, Denizli, Türkiye. She is currently pursuing the Ph.D. degree in computer engineering with Yıldız Technical University, İstanbul, Türkiye. She is a Research Assistant in computer engineering with Çanakkale Onsekiz Mart University. Her research interests include neural networks, machine learning, deep learning, and sign language recognition.

**SIRMA YAVUZ** (Member, IEEE) received the Ph.D. degree in computer engineering from Yıldız Technical University, İstanbul, Türkiye. She is currently a Professor with the Department of Computer Engineering, Yıldız Technical University. Her current research interests include artificial intelligence, neural networks, machines, and robotics.