Deep Learning-based Arabic Sign Recognition System for Automated Communication with Hearing Impaired Individuals

Runna Alghazo[1], Ghazanfar Latif[2*], Nazeeruddin Mohammad[2], Jaafar Alghazo[3], and Maura A. E. Pilotti[4]

[1] Inclusive Rehabilitation Sciences, Department of Education, Health, & Behavioral Studies (EHBS), University of North Dakota, Grand Forks, ND 58202, USA
[e-mail: runna.alghazo@und.edu]
[2] Computer Science Department, Prince Mohammad bin Fahd University
Al Khobar, 31952, Saudi Arabia
[e-mail: glatif@pmu.edu.sa, nmohammad@pmu.edu.sa]
[3] Software Engineering and Information Technology Management Departments, University of Minnesota, Crookston, USA
[e-mail: alghazo@crk.umn.edu]
[4] College of Sciences and Human Studies, Prince Mohammad bin Fahd University
Al Khobar, 31952, Saudi Arabia
[e-mail: mpilotti@pmu.edu.sa]
[*]Corresponding author: Ghazanfar Latif

Received March 7, 2024; revised August 29, 2024; accepted September 29, 2024; published November 30, 2024

Abstract

Arabic Sign Language (ArSL) is used by individuals who are hard of hearing or deaf in Arab countries, as well as others around the world who use it for religious purposes. for the need for automated systems to facilitate the learning and communication of ArSL is therefore significant. Such systems would allow people to learn Arabic Sign Language and use it to communicate among themselves and with the surrounding community. This paper presents the development of an automatic recognition system capable of accurately identifying Arabic signs through hand gestures. In this paper, two Residual Network (ResNet) Configurations, Version 1 (V1) and Version 2 (V2), are proposed and detailed. The proposed ResNet V1 achieved an average accuracy of 98.83%, while ResNet V2 achieved an average accuracy of 98.84%. The results described in this paper far exceed those reported in the extant literature. The high accuracy of the proposed system shows the potential for integrating the system with education tools and assistive technologies for people with special needs.

Keywords: Arabic Sign Language, Deep Neural Networks, ResNet, CNN, Automatic Recognition, Hearing Impaired.

http://doi.org/10.3837/tiis.2024.11.003 **ISSN : 1976-7277**

1. Introduction

Language is a medium through which we communicate meaning and establish relationships [1]. Spoken language involves two interconnected processes unique to the human species: speech perception and speech production [2]. Interconnectivity can be disrupted by hearing loss. According to the World Health Organization, 5.3% of the world's population suffers from disabling hearing loss, which is defined as a loss greater than 40 dB in adults' better hearing ear and greater than 30 dB in children's better hearing ear [3-4]. If the hearing loss is severe (70-90 dB), the listener cannot hear speech sounds and almost no other sound, whereas if the loss is profound (91+ dB), no sound at all is perceived [5]. As an audiological phenomenon, deafness, which may be present from birth or acquired, is generally defined [6-8] as a person's inability to understand speech using sound alone (with or without hearing aids or other devices that can serve as amplifiers). It results in the person’s inability to use hearing as the primary channel for receiving speech.

Hearing and deaf persons encounter a challenging communication barrier when attempting to interact with each other. Fortunately, human communication does not wholly rely on a single modality. Thus, alternative solutions to medical interventions [9] entail changing the channel upon which speech is perceived, such as switching from an auditory to a primarily visual mode. Visual means of perception include sign language, lip-reading, speech-reading, and reading and writing.

Within the ecosystem of the interchangeable roles of talker and receiver, sign language can be defined as a manual communication system comprised of unique signs that are articulated by the talker/signer to convey information and are to be apprehended by the receiver to be adequately processed. Because signs are produced by someone whose intent is to communicate meaning, they can be classified as symbols [10] that rely on two general systems of manual communication: (a) A sign language that consists of a set of manual configurations and gestures corresponding to content and function words. Like spoken languages, sign languages have their own combinatorial rules (grammar) and suffer from comparable national and regional variations (i.e., diglossia) [11]. (b) A finger spelling or manual alphabet, whereby the words of a language, such as Arabic, are spelled out manually. It consists of configurations of the human hand that correspond to letters of the alphabet. In everyday life, the utility of finger-spelling emerges in situations in which the word that the talker intends to communicate and the receiver wants to understand is not included in the manual vocabulary of either party (e.g., a novel word or a proper noun). Alternatively, it refers to situations in which the word is known, but noise or ambiguity in the human interaction requires further clarification of the intended meaning. Ambiguity may reach uncertainty and even utter confusion when the two interacting parties rely on different sign languages to communicate (American Sign Language, ASL, and Arabic Sign Language, ArSL), each unknown or little known to the other party. In the latter case, reliable cross-language translation is necessary for the intended meaning to be accurately conveyed to the receiver.

Interestingly, medical assessments and interventions are concerned primarily with the properties of the physical loss, such as its origin, degree, type, onset, and structural pathology, and much less with the communicative challenges it brings about and their implications (e.g., dependency, likely disruption of social relationships, etc.). Yet, the implications of deafness for the person who is experiencing it are noticeable [12]. It is important to note here that congenital deafness, through the early auditory deprivation that it produces, poses severe challenges to the intellectual, behavioral, cognitive, and social development of children. The onset of deafness is a relevant factor in shaping the severity of the quality-of-life outcomes of hearing loss. The effects of congenital deafness and deafness acquired in early childhood are very similar. Still, they differ markedly from those resulting from deafness acquired in late childhood or adulthood (e.g., occupational deafness and elderly deafness) [13]. For instance, the language deprivation of early-onset deafness can limit children's acquisition of social knowledge, thereby leading, among many of its potential outcomes, to social isolation, low self-esteem, and parental stress [14]. Deafness acquired in adulthood has different quality-of-life outcomes because everyday communication breaks down after a spoken language has already been learned and has been put to full use for quite some time. Compared with early-onset deaf persons, individuals suffering from late-onset deafness tend to be reluctant to change their usual means of communication, finding hearing loss an insurmountable obstacle. As such, they are more likely to experience embarrassment, loss of confidence, social isolation, and depression [15].

Although the experience of alienation from the larger hearing community may be felt, being deaf does not entail a view of oneself as handicapped or disabled. Instead, one is likely to see oneself as a member of the deaf community, a group of individuals who share, to some extent, a common language, life experiences, and a sense of cultural identity. Communication barriers exist not only for hearing and deaf individuals who are attempting to interact with each other but also for deaf individuals using different manual languages. Whether obstacles are conceptualized as hindrances or challenges that demand a problem-solving attitude has serious consequences on the lives of all parties involved. Not surprisingly, the concept of empowerment [12,16] has emerged as a useful tool for ameliorating views of communication challenges and for supporting a healthy problem-solving approach to the demands posed by social interactions in everyday life. The concept, rooted in the notion of complementarity, which promotes respect and acceptance of people as equals no matter their differences, comprises the dimensions of power-inside (e.g., acceptance and confidence in oneself), power-for (e.g., control over one's decision-making), and power-with (e.g., recognition of common goals and solidarity with members of one's community and of other communities).

Within the notion of each type of empowerment lies the recognition of the importance of technological devices that enrich communication channels in sensory modalities other than hearing. The issue that remains at the forefront of research in artificial intelligence and a matter of contention is the degree of accuracy of online translations of manual signs into their spoken or written intended counterparts or of translations in the opposite direction. In the real world, spontaneous and informal communications are much more likely to be affected by noise and ambiguity than in the rarified laboratory where intelligent human-computer interaction (HCI) models are devised and tested.

The goal of the research presented herein is to describe a bilingual alphabetical sign recognition system to aid communication (a) between deaf individuals who speak ArSL and wish to learn ASL (and vice versa) or (b) between deaf individuals who speak either of these languages and their hearing counterparts. The system may further aid ArSL or ASL users when words are unfamiliar, unknown, or merely ambiguous due to noise in the communication channel. The proposed system relies on the Arabic sign alphabet. It is recognized that the accurate identification of the constituents of meaningful utterances is important to learners who wish to gather mastery in a given language and, more broadly, to users who need to resolve ambiguities. In fact, due to the pervasive "diglossia" that characterizes different manifestations of ArSL, it is argued that an automated translation system between the Arabic alphabet and the English alphabet can be a useful tool for overcoming instances of miscommunication attributable to local variations in whole word signing.

The choice of the Arabic language, a Semitic tongue, is based on the fact that it is the habitual communication mode of approximately 260 million people who possess it as their first language [17]. Although Modern Standard Arabic (MSA) is the official language of Arab governments, widely used in formal education practices, Arab print, and broadcast media [18], it coexists with many national and regional dialects or vernaculars, likely to be as prevalent as MSA in particular local communities.

Although the estimated prevalence of the English language is much broader, the use of Arabic can be considered, at present, not only widespread but also growing [19]. Although the size of the deaf community in the Middle East is largely unknown, statistics involving specific countries suggest non-negligible numbers of members. ArSL may be the official sign language in Middle Eastern countries, but diglossia remains rampant. Thus, the proposed system is expected to have practical utility for a large array of constituencies whose needs cannot be left unanswered. In a globalized world, the physical distance between/among people can decrease only if improved communication is secured. Below we briefly review the findings of the extant literature on automated recognition systems whose goal is to serve the particular communication needs of the deaf community. The bilingual recognition system we propose is situated within the extant literature, demonstrating its potential utility.

The main contribution of this study is the development of an accurate ArSL recognition system using two novel revised versions of Residual Networks (ResNet V1 and ResNet V2). The results indicated higher ArSL recognition accuracy compared with the values reported in the extant literature. Furthermore, the study introduces an optimized network depth for ResNet models tailored to the specific challenges of recognizing ArSL, which is an issue that has received limited attention in prior research.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of the existing literature, highlighting the current challenges and approaches. Section 3 introduces the Convolutional Neural Network (CNN)-based recognition approach, detailing the architecture and components of the ResNet models used in this study. Section 4 describes the methodology, including the dataset, preprocessing steps, and the specific configurations of the ResNet models (V1 and V2). Section 5 presents the experimental results and discusses the performance of the proposed models compared to existing methods. Finally, Section 6 concludes the paper by summarizing the key findings and suggesting directions for future research.

2. Literature Review

Sign recognition can be realized using a sensor-based approach or an image-based approach [20]. In a sensor-based approach, the signer is required to wear a glove when making gestures to convey information. Multiple sensors read the hand gestures made. Evidence exists that sensor-based systems can be reliable and reasonably accurate [21]. However, they are often judged to be burdensome, intrusive, and unnatural by the user as he/she is required to wear a glove loaded with cables, sensors, and other technical materials. They also tend to be more expensive as they rely on the integration of software and hardware solutions. On the other hand, image-based systems overcome the burden of the signer's wearing any kind of gloves by using image processing techniques to recognize signs. In image-based systems, the extraction of particular features to be used by a learning algorithm for classification may be determined by the developer. Alternatively, features may be extracted in an automated manner from the input image through a series of algorithms. In the latter, also called a deep learning approach, the automated extraction of features is hierarchical. The features that are selected are those that most effectively define the input image.

Compared with research focusing on sign languages, research specifically devoted to automated recognition systems for ArSL (including finger-spelling and sign language of utterances) is of more recent development and thus less complete. Predictably, the preferable procedural approach is still a matter of debate. Feature extraction approaches have been said to lead to high classification accuracy rates if carefully engineered [22]. However, in some studies, deep learning approaches have been reported to yield better accuracy [23], whereas in other studies, the accuracy of feature extraction approaches and that of deep learning algorithms do not seem to be notably different [20].

In the research summarized below, the methodologies and findings of a selected collection of key studies on image-based recognition of Arabic alphabet signs are reviewed to situate our work in the proper niche of the extant literature. Although the sampled studies adopt the image-based approach for fingerspelling, whereby the signer executes the sign of each letter separately, they differ in a variety of ways, including the datasets, techniques, and algorithms modeling the proposed applications. Yet, the goal is the same. Namely, the development and assessment of procedural and computational solutions aimed at finding the optimal system for accurately recognizing hand configurations, including position, orientation, and conceivably movements. Regardless of the assortment of modes of operation, test findings are consistently encouraging, yielding high recognition rates across the board. For instance, Elsayed and Fathy used a deep Convolutional Network (CNN) approach for feature extraction and recognition of ArSL [24]. They combined the power of Web semantics with deep CNNs. Their method achieved 88.87% accuracy on the ArSL dataset. Similarly, Saleh and Issa used deep CNNs to enhance the recognition accuracy of 32 ArSL gestures [25]. They chose pre-trained VGG-16 and ResNet-152 models and then fine-tuned these models by retraining them after adding additional layers. The fine-tuned ResNet model achieved the highest accuracy of 99%. Nurnoby and colleagues also used pre-trained CNN models on a huge dataset and fine-tuned them on an ArSL dataset [26]. Their aim was to improve recognition accuracy when the signs have complex backgrounds. Their approach achieved 94.33% accuracy on an ArSL dataset. Taken together, the studies reviewed here lead to the conclusion that the CNN-based recognition approach is particularly advantageous. Of course, innovative solutions that improve over existing ones remain to be sought.

3. Convolutional Neural Network-based Recognition

A convolutional neural network (CNN) is a class of deep neural networks that are commonly used for image classification. The input data for a CNN are images $x_1, x_2, x_3, \dots, x_n$, which are fed into a convolutional layer. Image x_n can be formally described as $M \times M \times C$, where M refers to the height and width of the image. Thus, $M \times M$ is the number of pixels or the resolution of an image. C is the number of channels in the image. The number of channels varies depending on the image type. For greyscale images $C = 1$, whereas, for colored images (RGB) $C = 3$. Generally, a CNN architecture consists of convolutional layers, pooling (subsampling) layers, and fully connected layers [27].

A convolutional layer consists of a set of learnable filters or kernels that extract different feature maps, which are later used during the classification phase. Each kernel has the dimensions of $N \times N \times R$, where $N \times N$ is the height and width of the kernel, and R refers to the number of channels, which is the same as or less than the channels C in the convolved images [27]. In a convolutional layer, each filter (kernel) can extract specific features by convolving the input image, thereby, producing k feature maps of the size $M - N + 1$, as shown in **Fig. 1**. Each feature map is then down-sampled (pooled) using mean or max pooling over $q \times q$, where q is the number of strides, which typically ranges between 2 to 5 for large inputs, as shown in **Fig. 2**. Finally, succeeding all convolutional and pooling is a number of fully connected layers.

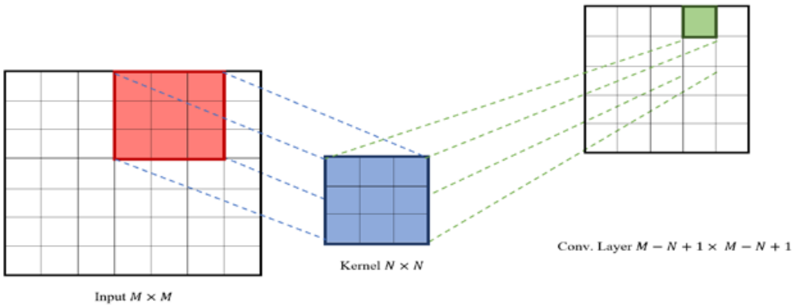

Fig. 1. Mapping of the convolutional kernel filter to the input Image

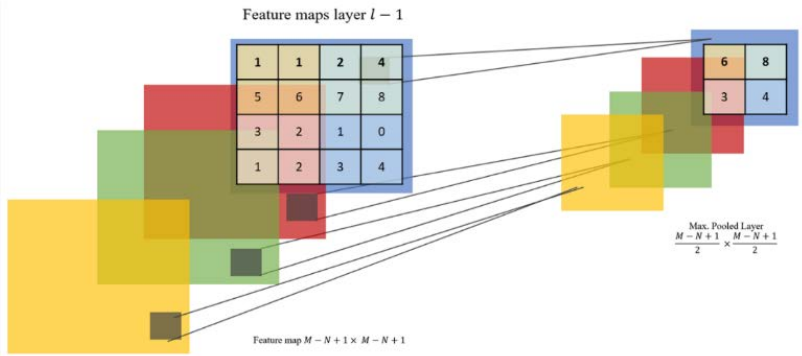
Feature maps layer $l - 1$
1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
6 8
3 4
Max. Pooled Layer $\frac{M-N+1}{2} \times \frac{M-N+1}{2}$
Feature map $M - N + 1 \times M - N + 1$

Fig. 2. Sample of the max pooled layer mapping to the feature map

Convolutional Layers act as feature extractors that unearth prominent features, such as edges, corners, and endpoints, from input images to compute pre-nonlinearity input for some units. Consider a convolutional layer l that receives an input image of size $M \times M$ and convolves it with a set of filters k with $N \times N$ dimensions. It produces k-feature maps with an output of size $(M - N + 1) \times (M - N + 1)$, illustrating a single convolutional layer that produces k-feature maps. Furthermore, extracted feature maps from layer l may be passed to other convolutional layers to extract higher-level features from the input image. An activation function is applied following each convolutional layer.

Pooling Layers or subsampling layers may follow each convolutional layer in a CNN. Pooling layers down-sample each convolutional layer output, thereby reducing its spatial height and weight dimension. The main purpose of this process is to reduce the parameters learned by the network, thus cutting the computations performed as well as lowering the resolution of the images. Namely, the desired outcome is to reduce the precision of the translation effect and generalize the feature maps (1, 5). There are several methods to perform pooling, such as max pooling and average pooling. Nonetheless, pooling in general works by dividing the resultant feature map from convolutional layers into several regions and subsampling each region individually.

Fully connected Layers are used for the output features after several convolutional and pooling layers. The primary function of this process is to take all units (i.e., neurons) in the previous layer (pooling, convolutional, or fully connected) and connect them to every single neuron in the next layer. A fully connected layer is used to classify the high-level features produced by the convolutional and pooling layers.

4. Methodology

This paper illustrates experiments conducted using our own ArSL dataset (ArSL2018) [28]. This dataset is available to other researchers to check the performance of newly proposed classification algorithms and methodologies. We have previously applied CNN and Random Forest (using geometric features) to the dataset. CNN models use a common activation function, Rectified Linear Units (ReLU), which returns the result value of the convolutional layer if it is positive. Otherwise, it returns zero. ReLU is used to increase the non-linear properties of the network's decision function without affecting the convolutional layer. In this paper, we are applying two versions of residual networks (ResNets) with proposed optimized network depths. The details of the dataset and proposed ResNet versions along with relevant background are discussed in the following subsections.

4.1 Dataset

The ArSL2018 is a comprehensive, fully labeled ArSL images dataset that was developed at Prince Mohammad Bin Fahd University (PMU). After the items of the ArSL2018 dataset were collected, labeled, and systematized, they were made publicly available [28]. Thus, the dataset is now accessible to researchers to benefit communicative exchanges involving deaf and hard-of-hearing individuals. To our knowledge, the ArSL2018 dataset is unique because it is the first comprehensive dataset of ArSL. As such, it is particularly suited to test the accuracy of classification and recognition of various applications as well as to develop prototypes useful to the deaf community.

The ArSL2018 dataset is composed of 54049 images, each with 48 x 48 dimensions. **Fig. 3** displays sample images of the ArSL alphabet with labels and numbers for the available images. The dataset can be used 'as-is' and may be augmented with additional variants from a second version of the dataset. Limitations of the current ArSL2018 dataset include the following issues: 1) the number of lighting and noise variations of the current sample is narrow, and 2) a limited number of participants provided the samples (*n* = 40). The developers are in the process of addressing these limitations, which are viewed as minor, in the next realization of the dataset.

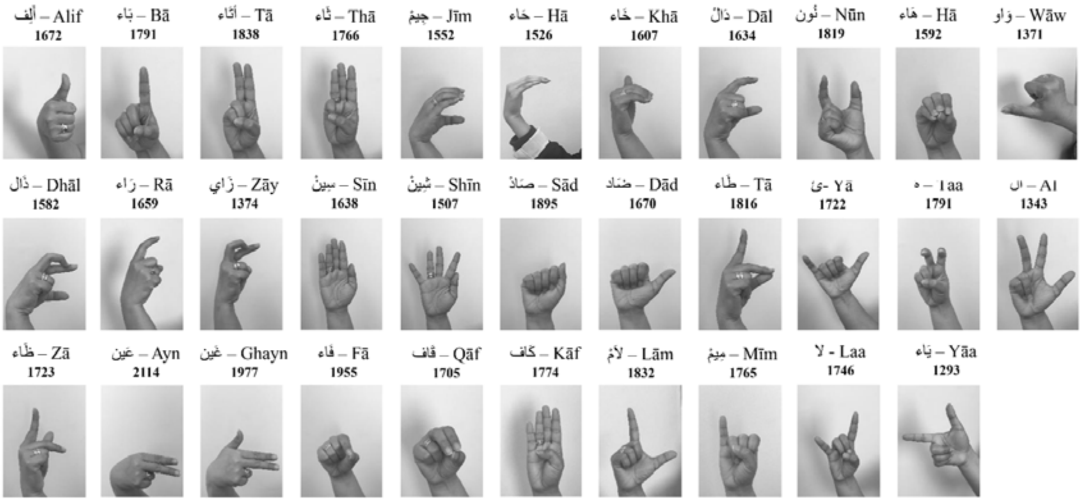

Fig. 3. Visualization of the Arabic Alphabet of the ArSL

4.2 Preprocessing

The ArSL dataset was generated locally at PMU with the help of volunteers of different ages. The distance between the camera and the volunteers was approximately one meter. In the images, light intensity, background, and camera angle varied. The dataset was preprocessed using Matlab software to make it more appropriate for image classification using machine learning approaches. Images were converted to grayscale since color information is not useful for sign recognition. All images were then resized to a fixed dimension of 64 × 64. Features were intended to be extracted from these preprocessed fixed-dimensional images. Additional preprocessing techniques, such as normalization and contrast adjustment, were applied to enhance the dataset's robustness.

4.3 Classification using Enhanced Residual Neural Networks (ResNets)

CNN can assimilate different image features and classifiers in an end-to-end system with minimal preprocessing. These characteristics have had a great impact on image processing research. Deep CNN (DCNN) with several additional layers (i.e., more network depth) can help extract better features. Nonetheless, deep networks are affected by degradation issues once they reach a convergence point. It was noted that the accuracy of deep networks gets saturated and quickly degrades as the number of layers is increased [29]. The degradation cannot be attributed to the overfitting of models, as the training accuracy also degrades. The problem of training deep networks has been addressed by residual blocks being added to the network [29-30].

In our research, a Residual Neural Network (ResNet) was used because it is a different class of artificial neural networks (ANN) that utilizes special additional connections called skip connections. These connections directly join one layer's output with the following layer's output. Empirical research suggests that shortcut connections can produce a smoother optimization landscape, and gradient decays sub-linearly instead of exponentially as in standard ANNs. Thus, shortcut connections relieve the vanishing gradient descent problem. Further benefits of shortcut connections are evidenced in [30].

The residual function (Ƒ) for a simple block of ResNet shown in **Fig. 4** is given by $Ƒ = W_2 A(W_1 x)$. Here, A is the activation function (such as ReLU), and x is the input to the layers. The output (y) of the block is $Ƒ(x) + \mathcal{P}(x)$. It can be generalized as in (1).

$$y = Ƒ(x, \{W_i\}) + x \qquad (1)$$

In case the dimensions of Ƒ and x are not the same, a linear projection of skip connections with W_s can be used to eliminate the dimensions' mismatch. The resulting output block can now be expressed as in (2).

$$y = Ƒ(x, \{W_i\}) + W_s x \qquad (2)$$

The residual function $F(x, \{W_i\})$ is the mapping that will be learned, and it represents several convolutional layers.

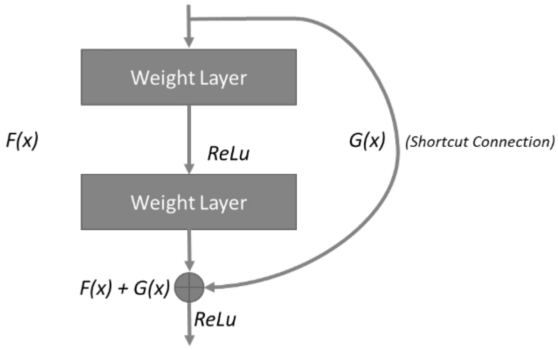

Fig. 4. Residual Network Block Structure

ArSL classification is performed using a modified residual neural network model with varying input parameters and a varying set of layers. **Fig. 5** and **Fig. 6** show the two modified ResNet models. The shortcut identity connections are integrated within every block of the 3×3 layers in ResNet model 1. As indicated in the figures, identity mapping is utilized for all the shortcut connections. In cases where the output and the input dimensions are different, a projection shortcut (using a 1×1 convolution layer) is used. Batch normalization and nonlinear activation are used for the shortcut connection to avoid degradation and vanishing gradient problems. The residual block in (3) controls the depth of the convolution layers in the ResNet model 1.

$$depth\_V1 = N * 6 + 2 \qquad (3)$$

The number of residual blocks is designated by N, e.g., *N=4 for ResNet 26.* In the model, j is the number of loops (1: N), and i is the number of stages in the model (both shown in **Fig. 5**).

Bottleneck connections with filter size are calculated based on the increase of the block size in shortcut connections by multiplying Block size by 9 as presented in **Fig. 6**. Convolutional layers of size 1×1, 3×3, and 1×1, are the three-layer present in the residual function block. The input dimensions are increased and decreased using the 1×1 layers. The smaller dimensions face a bottleneck from the 3×3 layer. The residual block in (4) controls the depth of the convolution layers in the ResNet model 2

$$depth\_V2 = N * 9 + 2 \qquad (4)$$

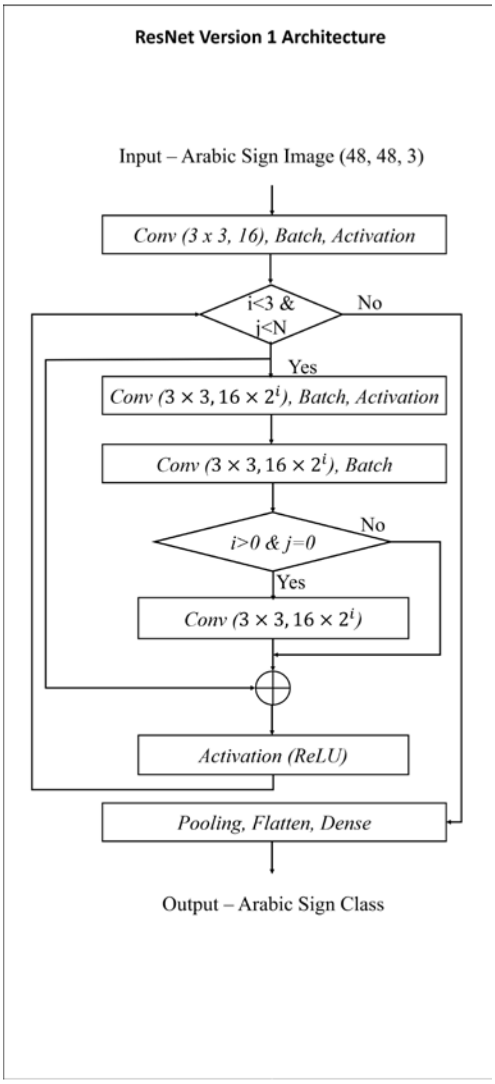
ResNet Version 1 Architecture
Input – Arabic Sign Image (48, 48, 3)
Conv (3 x 3, 16), Batch, Activation
i<3 & j<N
No
Yes
Conv (3 × 3, 16 × 2^i), Batch, Activation
Conv (3 × 3, 16 × 2^i), Batch
i>0 & j=0
No
Yes
Conv (3 × 3, 16 × 2^i)
Activation (ReLU)
Pooling, Flatten, Dense
Output – Arabic Sign Class

Fig. 5. The architecture of the ResNet model 1

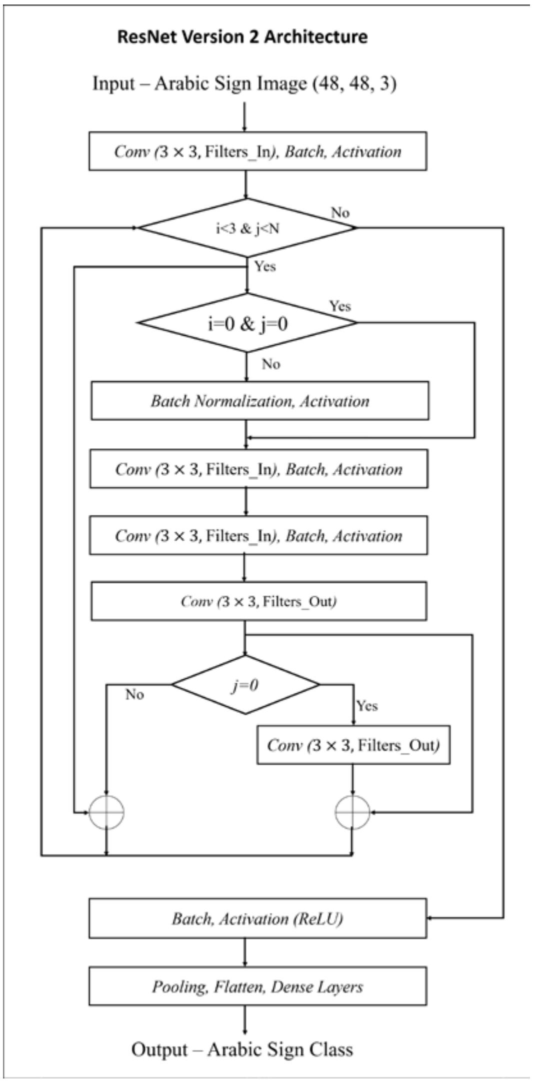
ResNet Version 2 Architecture
Input – Arabic Sign Image (48, 48, 3)
Conv (3 × 3, Filters_In), Batch, Activation
i<3 & j<N
No
Yes
i=0 & j=0
Yes
No
Batch Normalization, Activation
Conv (3 × 3, Filters_In), Batch, Activation
Conv (3 × 3, Filters_In), Batch, Activation
Conv (3 × 3, Filters_Out)
j=0
No
Yes
Conv (3 × 3, Filters_Out)
Batch, Activation (ReLU)
Pooling, Flatten, Dense Layers
Output – Arabic Sign Class

Fig. 6. The architecture of the ResNet model 2

5. Results and Discussions

The proposed approaches were implemented in Python programming language using TensorFlow libraries. The Python programs were executed on a six-core 3.70GHz CPU with 32 GB RAM and Nvidia GeForce GTX-1080 GPU with 2560 CUDA cores. The input for the proposed models was a pre-processed dataset, which was split into train and test sets with an 80-20 ratio. The training set was further split into two parts with an 80-20 ratio. In it, 20% of the training set was used for validation.

The experiments were performed on the described dataset using the ResNet V1 and ResNet V2 with various proposed depths, various parameters, and varying blocks. As detailed in **Table 2**, varying the blocks from 1 to 3, and then to 5 resulted in a depth of 6, 20, and 32 for ResNet V1, respectively. Also shown in **Table 2** is that varying the blocks from 1 to 3, and then to 5, resulted in depths of 11, 29, and 47 for ResNet V2, respectively. To elaborate, the depths of the two versions of ResNet (V1 and V2) were selected based on empirical testing. Initially, various depths were tested, ranging from shallow networks to more complex architectures, to identify the optimal balance between model performance and computational efficiency.

Fig. 7 shows the results of the training loss of the different ResNet depths for both V1 and V2. The curves shown in the figure are all converging close to zero, thereby showing that all models were learning well. To verify whether they were not overfitting and whether generalization was a viable outcome, validation loss was also calculated and plotted.

Fig. 8 shows the validation loss which was calculated exactly in the same way the training loss was calculated except that it was used to update the weights, whereas training loss was not used to update the weights. If the results obtained in **Fig. 7** and **Fig. 8** are compared, it can be observed that the training loss and the validation loss were stabilized after epoch 80, and they reached almost the same value. This outcome underscores that the model had a nearly perfect fitting, neither overfitting nor underfitting. Thus, the model demonstrated excellent generalization capabilities in predicting and classifying new data.

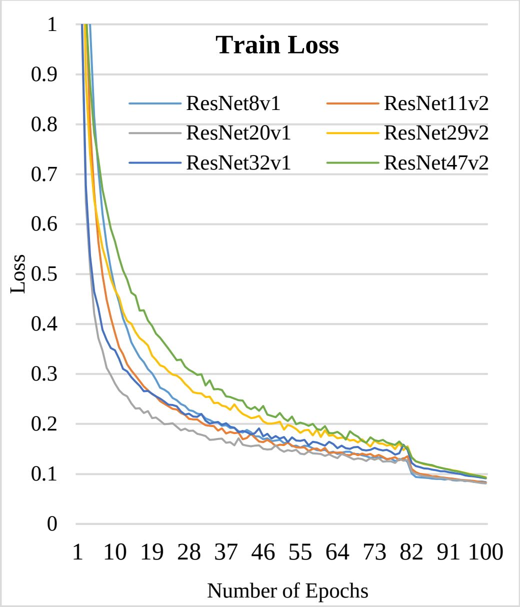

Fig. 7. Comparison of training loss

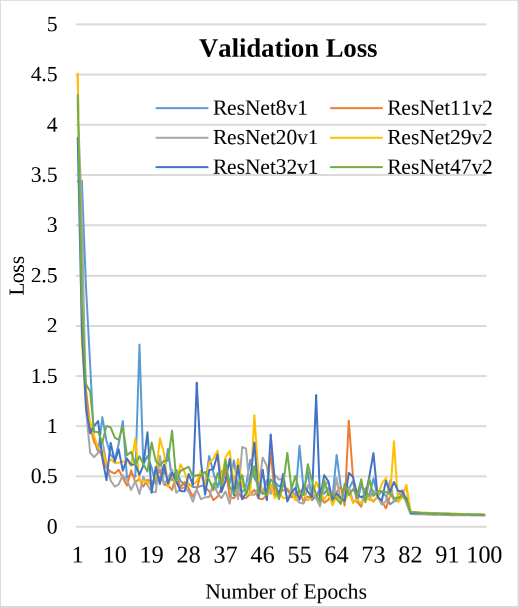

Fig. 8. Comparison of validation loss

Fig. 9 plots the training accuracy of both versions of the ResNet architectures proposed in this research. It is important to note that the training accuracy for both ResNet V1 and V2 with various depths reached a stable saturated state, thereby ensuring a good training accuracy as well as indicating that the model could generalize and classify new data well.

Fig. 10 shows the validation accuracy for the proposed ResNet Versions V1 and V2 with different depths. If the validation obtained in **Fig. 10** is compared with the training obtained in **Fig. 9**, it can be seen that both the validation and training had reached almost similar values. Namely, the gap between training and validation was minimal, thereby illustrating a good fit.

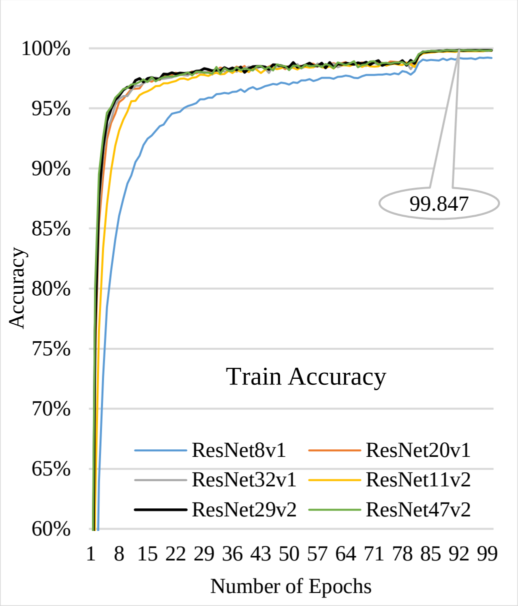

Fig. 9. Comparison of training accuracy

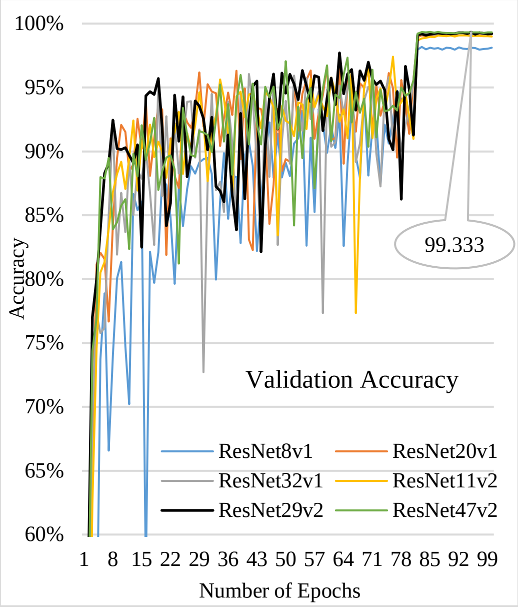

Fig. 10. Comparison of validation accuracy

Table 1 summarizes the comparison of the results obtained by the two proposed ResNet Versions V1 and V2 with different epochs, blocks, depths, and parameters. The V1 had depths of 6, 20, and 32, whereas V2 had depths of 11, 29, and 47. As explained earlier, depths cannot be the same because the equation for depth is different between ResNet versions (#blocks × 6 + 2 for V1 and #block × 9 + 2 for V2). As illustrated in **Table 1**, the best accuracy for ResNet V1 was 99.83%, which was achieved at depth 20 with 100 Epochs. The corresponding validation was 99.18% with a training loss and validation loss of 0.081 and 0.113, respectively. The best training accuracy for ResNet V2 was 99.84%, which was achieved at depth 29 with 100 epochs. The corresponding validation accuracy was 99.33% with training loss and validation loss of 0.094 and 0.12, respectively. One important point that can be observed here is that there is not much difference between the accuracy values of ResNet V1 and V2. Namely, for this dataset, the additional complexity of ResNet V2 was not beneficial to its performance. Similarly, increasing the number of blocks from 3 to 5 decreased the performance of both ResNet versions.

Experiments were also performed using the feature-based method where the feature engineering process consisted of manually extracting the features from the images that would produce the best results. Informed by the literature related to ArSL alphabet recognition, in this research, more consideration was given to local structural features rather than global structural features. For ArSL classification, around 40 geometric features were selected from the pre-processed ArSL images. Some of these features included (a) the area and the total circumference of images based on the total white pixels; (b) the maximum/minimum height and width of consecutive connected white pixels; (c) the maximum/minimum vertical and horizontal starting and ending locations for both axes; (d) the proportion of white pixels in the upper and bottom halves of images; and (e) centroids, the number of edges, and the roundness of ArSL images.

The precision and recall metrics were also calculated taking the configuration of the ResNet V1 and V2 that produced the best test accuracy shown in **Table 2**. A 99.64% precision score and a 98.88% recall score were achieved by ResNet V1 at a depth of 20 and with 100 epochs.

A 99.95% precision score and 99.33% recall score were achieved by ResNet V2 at a depth of 29 and with 100 epochs. Overall, the results obtained in this work outperform the outcomes reported in the extant literature.

Table 1. Comparison of ResNet V1 and V2 train/validation accuracies and loss with different epochs, and network depths

ResNet Version	ResNet Blocks	Network Depth	Total Parameters	Epochs	Train Accuracy	Validation Accuracy	Training Loss	Validation Loss
V1	1	1x6+2=8	80,096	25	95.20%	84.16%	0.247	0.567
				50	96.98%	89.12%	0.168	0.403
				75	97.81%	90.62%	0.133	0.353
				100	99.21%	98.11%	0.084	0.121
V1	3	3x6+2=20	275,872	25	97.88%	93.24%	0.195	0.338
				50	98.53%	89.38%	0.149	0.469
				75	98.91%	96.14%	0.124	0.222
				100	**99.83%**	**99.18%**	**0.081**	**0.113**
V1	5	5x6+2=32	471,648	25	97.74%	90.92%	0.236	0.454
				50	98.53%	93.95%	0.171	0.309
				75	98.69%	95.81%	0.147	0.258
				100	99.80%	99.16%	0.091	0.118
V2	1	1x9+2=11	304,832	25	97.39%	88.24%	0.229	0.515
				50	98.42%	92.40%	0.159	0.353
				75	98.66%	94.87%	0.134	0.256
				100	99.79%	98.98%	0.083	0.112
V2	3	3x9+2=29	854,656	25	97.94%	94.28%	0.297	0.427
				50	98.44%	94.56%	0.205	0.339
				75	98.72%	90.97%	0.161	0.429
				100	**99.84%**	**99.33%**	**0.094**	**0.120**
V2	5	5x9+2=47	1,404,480	25	98.00%	93.41%	0.328	0.458
				50	98.19%	97.06%	0.222	0.275
				75	98.70%	93.22%	0.168	0.353
				100	99.83%	99.32%	0.092	0.114

The results of the present research support the use of CNNs for the recognition of Arabic Hand Sign Language. Two Residual Network configurations were put forth, labeled Version 1 and Version 2, and described thoroughly. The use of ResNet V1 and ResNet V2 for the recognition of ArSL achieved a test accuracy of 99.83% and 99.84%, respectively. Additionally, various metrics were presented to show that the proposed models neither overfit nor underfit and that they generalize well when the task is to predict the classification of new data. Comparisons of the results obtained in this paper with the results reported in the extant literature show that the proposed models outperform and produce better results than those obtained by earlier models (see **Table 2**).

Table 2. Comparison of the test accuracy of the proposed models with that of other models using the same dataset

Proposed Method	Accuracy	Precision	Recall
ResNet V1 – Depth 20 (proposed)	99.80%	99.64%	98.88%
ResNet V2 - Depth 29 (proposed)	**99.83%**	**99.95%**	**99.21%**
Geometric Features + RF (proposed)	92.15%	96.24%	96.92%
Semantic Deep Learning (2020) [24]	88.87%	NA	NA

CNN - VGG16 (2020) [25]	89.89%	NA	NA
ResNet152 (2020) [26]	99%	NA	NA
CNN Model 1 (2020) [27]	95.90%	NA	NA
CNN Model 2 (2020) [27]	97.60%	NA	NA
Vision Transformers based Transfer Learning (2023) [31]	98%	NA	NA
modified inception V3 (2023) [32]	97.4%	NA	NA
Vision Transformer [33]	99.3%	NA	NA

6. Conclusion

Systems that automatically recognize ArSL are necessary for developing state-of-the-art systems that can assist individuals who are hearing impaired or individuals who work and communicate with those who are hearing impaired. Our work, which combines the contributions of a multidisciplinary team of scientists, aims to develop an automatic recognition system for ArSL with high accuracy, using convolutional neural networks, such as Residual Networks. In this paper, two Residual Network Versions V1 and V2 were proposed, each with a detailed configuration. Experiments using our ArSL dataset were performed. An accuracy of 99.83% and 98.84% was achieved for ResNet V1 and ResNet V2, respectively. Interestingly, the selected models were shown to be neither overfitting nor underfitting, thereby indicating that they generalized well to new data. Further work in this field will include expanding our research to include sign language from a variety of languages, especially languages that have been neglected by other researchers. Also, future work will include the use of Deep Learning Networks for the development of various systems to assist individuals with disabilities, such as those who are visually impaired. Future research could focus on several key areas to build upon our findings. Enhancing the model to address any remaining challenges, such as misclassifications in similar hand gestures, could further improve its robustness. Additionally, expanding the dataset to include more diverse signers and different dialects would enhance the model's generalizability. Finally, integrating our system into practical applications, such as real-time translation tools or educational platforms, represents a promising direction for future work.

In summary, this study presents a novel approach to Arabic Sign Language recognition using optimized ResNet architectures. Our findings demonstrate significant improvements in recognition accuracy compared to existing methods, highlighting the potential of deep learning models in this domain.

The broader implications of our work extend beyond the specific case of Arabic Sign Language. By advancing the capabilities of sign language recognition systems, our research contributes to the development of more inclusive technologies that can bridge communication gaps for the deaf community. This work also sets the stage for further exploration of deep learning applications in other specialized language and gesture recognition tasks.

References

[1] K. A. Noels, T. Yashima, and R. Zhang, Language, identity, and intercultural communication, 2nd Edition, The Routledge Handbook of Language and Intercultural Communication, Routledge eBooks, pp.55-69, May. 2020. Article (CrossRef Link)

[2] K. J. Forseth, G. Hickok, P. S. Rollo, and N. Tandon, “Language prediction mechanisms in human auditory cortex,” *Nature Communications*, vol.11, no.1, Oct. 2020. Article (CrossRef Link)

[3] A. C. Davis and H. J. Hoffman, “Hearing loss: rising prevalence and impact,” *Bulletin of The World Health Organization*, vol.97, no.10, pp.646-646A, Oct. 2019. Article (CrossRef Link)

[4] R. Kushalnagar, “Deafness and Hearing Loss,” *Web Accessibility, Human-computer interaction series*, pp.35-47, Jun. 2019. Article (CrossRef Link)

[5] J. Andrews, P. Shaw, and G. Lomas, Deaf and Hard of Hearing Students, Handbook of special education, Taylor & Francis, Routledge, pp.241-254, 2011. Article (CrossRef Link)

[6] M. D. Felicite, “Glimpse in the World of Deaf People: Deafness and Deaf Education,” *GPH-International Journal of Social Science and Humanities Research*, vol.4, no.01, pp.12-30, 2021. Article (CrossRef Link)

[7] F. Sidera, G. Morgan, and E. Serrat, “Understanding Pretend Emotions in Children Who Are Deaf and Hard of Hearing,” *Journal of Deaf Studies and Deaf Education*, vol.25, no.2, pp.141-152, Apr. 2020. Article (CrossRef Link)

[8] S. Shave, C. Botti, and K. Kwong, “Congenital Sensorineural Hearing Loss,” *Pediatric Clinics of North America*, vol.69, no.2, pp.221-234, Apr. 2022. Article (CrossRef Link)

[9] M. L. Hall and S. Dills, “The Limits of “Communication Mode” as a Construct,” *Journal of Deaf Studies and Deaf Education*, vol.25, no.4, pp.383-397, Oct. 2020. Article (CrossRef Link)

[10] S. Baowidan, “Improving realism in automated fingerspelling of American sign language,” *Machine Translation*, vol.35, no.3, pp.387-404, Sep. 2021. Article (CrossRef Link)

[11] A. Kusters, “International Sign and American Sign Language as Different Types of Global Deaf Lingua Francas,” *Sign Language Studies*, vol.21, no.4, pp.391-426, 2021. Article (CrossRef Link)

[12] T. Shalev, S. Schwartz, P. Miller, and B.-S. Hadad, “Do deaf individuals have better visual skills in the periphery? Evidence from processing facial attributes,” *Visual Cognition*, vol.28, no.3, pp.205-217, 2020. Article (CrossRef Link)

[13] J. Dammeyer, K. Crowe, M. Marschark, and M. Rosica, “Work and Employment Characteristics of Deaf and Hard-of-Hearing Adults,” *Journal of Deaf Studies and Deaf Education*, vol.24, no.4, pp.386-395, Oct. 2019. Article (CrossRef Link)

[14] F. Chang, H. X. Wu, B. H-H. Ching, X. Li, and T. T. Chen, “Behavior Problems in Deaf/Hard-of-Hearing Children: Contributions of Parental Stress and Parenting Styles,” *Journal of Developmental and Physical Disabilities*, vol.35, no.4, pp.607-630, Aug. 2023. Article (CrossRef Link)

[15] G. Movallali, Z. M. usavi, and E. Hakimi- Rad, “Feeling of Loneliness in Deaf Adolescents: the Effect of An Online Life Skills Program,” *European Journal of Social Science Education and Research*, vol.7, no.2, pp.1-14, Aug. 2020. Article (CrossRef Link)

[16] G. A. M. De Clerck, “Meeting Global Deaf Peers, Visiting Ideal Deaf Places: Deaf Ways of Education Leading to Empowerment, An Exploratory Case Study,” *American Annals of the Deaf*, vol.152, no.1, pp.5-19, 2007. Article (CrossRef Link)

[17] Y. A. A. Al-Nahdi and S. Zhao, “Learning Arabic language in China: Investigation on instrumental and integrative motivations of Chinese Arabic learners,” *Technium Social Sciences Journal*, vol.27, no.1, pp.767-797, Jan. 2022. Article (CrossRef Link)

[18] T. D. Wolsey, I. M. Karkouti, E. H. Hiebert, D. A. El Seoud, H. Abadzi, and F. Abdelkhalek, “Texts for reading instruction and the most common words in modern standard Arabic: an investigation,” *Reading and Writing*, vol.36, no.7, pp.1567-1587, Sep. 2023. Article (CrossRef Link)

[19] A. Gebril and H. Taha-Thomure, Assessing Arabic, The Companion to Language Assessment, vol. IV, pp.1779-1789, Nov. 2013. Article (CrossRef Link)

[20] M. Mohandes, S. Aliyu, and M. Deriche, “Arabic sign language recognition using the leap motion controller,” in *Proc. of 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)*, pp.960-965, Jun. 2014. Article (CrossRef Link)

[21] M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih, and M. M. bin Lakulu, “A Review on Systems-Based Sensory Gloves for Sign Language Recognition State of the Art between 2007 and 2017,” *Sensors*, vol.18, no.7, Jul. 2018. Article (CrossRef Link)

[22] T. Shanableh, K. Assaleh, and M. Al-Rousan, "Spatio-Temporal Feature-Extraction Techniques for Isolated Gesture Recognition in Arabic Sign Language," *IEEE transactions on systems, man, and cybernetics, Part B (Cybernetics)*, vol.37, no.3, pp.641-650, Jun. 2007. Article (CrossRef Link)

[23] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol.28, no.12, pp.3941-3951, Dec. 2017. Article (CrossRef Link)

[24] E. K. Elsayed and D. R. Fathy, "Sign Language Semantic Translation System using Ontology and Deep Learning," *International Journal of Advanced Computer Science and Applications*, vol.11, no.1, 2020. Article (CrossRef Link)

[25] Y. Saleh and G. F. Issa, "Arabic Sign Language Recognition through Deep Neural Networks Fine-Tuning," *International journal of online and biomedical engineering*, vol.16, no.05, pp.71-83, May 2020. Article (CrossRef Link)

[26] M. F. Nurnoby, E.-S. M. El-Alfy, and H. Luqman, "Evaluation of CNN Models with Transfer Learning for Recognition of Sign Language Alphabets with Complex Background," in *Proc. of 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*, pp.1-6, Dec. 2020. Article (CrossRef Link)

[27] G. Latif, N. Mohammad, R. AlKhalaf, R. AlKhalaf, J. Alghazo, and M. Khan, "An Automatic Arabic Sign Language Recognition System based on Deep CNN: An Assistive System for the Deaf and Hard of Hearing," *International Journal of Computing and Digital Systems*, vol.9, no.4, pp.715-724, Jul. 2020. Article (CrossRef Link)

[28] G. Latif, N. Mohammad, J. Alghazo, R. AlKhalaf, and R. AlKhalaf, "ArASL: Arabic Alphabets Sign Language Dataset," *Data in Brief*, vol.23, Apr. 2019. Article (CrossRef Link)

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016. Article (CrossRef Link)

[30] T. Liu, M. Chen, M. Zhou, S. S. Du, E. Zhou, and T. Zhao, "Towards Understanding the Importance of Shortcut Connections in Residual Networks," in *Proc. of the 33rd International Conference on Neural Information Processing Systems*, pp.7892-7902, 2019. Article (CrossRef Link)

[31] N. M. Alharthi and S. M. Alzahrani, "Vision Transformers and Transfer Learning Approaches for Arabic Sign Language Recognition," *Applied sciences*, vol.13, no.21, Oct. 2023. Article (CrossRef Link)

[32] B. Karsh, R. H. Laskar, and R. K. Karsh, "mIV3Net: modified inception V3 network for hand gesture recognition," *Multimedia Tools and Applications*, vol.83, no.4, pp.10587-10613, Jan. 2024. Article (CrossRef Link)

[33] A. F. Alnabih, A. Y. Maghari, "Arabic sign language letters recognition using Vision Transformer," *Multimedia Tools and Applications*, vol.83, no.34, pp.81725-81739, Oct. 2024. Article (CrossRef Link)

Runna Alghazo is an educational researcher and rehabilitation counselor, currently serving as an Assistant Professor in the College of Education and Human Development's Inclusive Rehabilitation Sciences program at the University of North Dakota, USA. She holds a Ph.D. in Rehabilitation Counseling and Administration from Southern Illinois University. Alghazo's research includes exploring applications of Artificial Intelligence to support students with disabilities, examining psychological and systemic factors influencing students' academic success, and developing inclusive teaching models based on Universal Design principles. Her work is dedicated to advancing accessibility and promoting inclusive practices in education.

Ghazanfar Latif is research coordinator (Deanship of Graduate Studies and Research) and Acting Director, Center for Artificial Intelligence at Prince Mohammad bin Fahd University, Saudi Arabia. He has been awarded 13 US patents and has published more than 100 articles in highly reputed journals and conferences. He holds Ph.D. degree in Artificial Intelligence (Computer Science) and was also post-doctoral fellow at University of Quebec, Canada. He earned his MS degree in Computer Science from King Fahd University of Petroleum and Minerals, Saudi Arabia in 2014 and BS degree in Computer Science from FAST National University of Computer and Emerging Sciences in 2010 by remaining on Dean’s honor list. Throughout his educational carrier, he got a number of achievements like a full scholarship for FSc, BS-CS, and MS-CS and a Gold Medal in Ph.D. He worked as an Instructor at Prince Mohammad bin Fahd University, Saudi Arabia for 3 years in CS Department and has 2 years of industry work experience. His research interests include Image Processing, Artificial Intelligence, Neural Networks, and Medical Image Processing.

Nazeeruddin Mohammad is a cybersecurity professional and researcher with 20+ years of combined experience in academic and industrial research, Infrastructure design/management, university-level teaching and proto-type development in simulation environments as well as in real world test-bed environments. He has extensive experience in designing, advising, implementing, and troubleshooting Information Technology (IT) systems/projects. Some of the projects he worked on: perimeter security design (firewall and proxy server implementation), disaster recovery (DR) site design and implementation, Storage Area Network (SAN) implementation, datacenter virtualization, web service platform implementation, network services design, and network/security policy design. His experience is across various major IT areas including computer networks, cloud computingand cybersecurity. He has certifications in Ethical hacking, CCNA (Routing and Switching) and Cyber Operations.

Jaafar Alghazo joined the University of Minnesota Crookston on August 12, 2024, as an Associate Professor with a dual appointment in the Software Engineering and Information Technology Management Programs. Before this, he was a Research Associate Professor under the Artificial Intelligence Research Initiative at the University of North Dakota's College of Engineering and Mines. Alghazo earned his Ph.D. and M.S. in Computer Engineering from Southern Illinois University at Carbondale (SIUC), Illinois, after completing his B.S. in Electrical Engineering from Mutah University, Jordan. Before joining UND, he was an Associate Professor in the Electrical and Computer Engineering Department at the Virginia Military Institute in Lexington. Alghazo also has extensive international experience, having spent 14 years at Prince Mohammad Bin Fahd University in Saudi Arabia, where he began as an Assistant Professor and later became an Associate Professor. He gained valuable administrative, educational, and research experience during his tenure. Additionally, he has taught at the University of Central Florida for two years and the American University in Dubai for another two years. Alghazo’s research interests are focused on Artificial Intelligence. He has authored or co-authored over 40 peer-reviewed journal papers, more than 30 conference papers, several book chapters, and a patent.

Maura A. E. Pilotti is a cognitive psychologist whose research interests include learning and memory processes across the lifespan. Currently, her research focuses on the interrelations of memory, language, and emotion. She received her Ph.D. in Cognitive Psychology from the City University of New York (USA). Institutional affiliation: Prince Mohammad Bin Fahd University (PMU).