IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 September 2022, accepted 25 September 2022, date of publication 28 September 2022,
date of current version 7 October 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3210543*

RESEARCH ARTICLE

# Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation

**B. NATARAJAN[1], E. RAJALAKSHMI[1], R. ELAKKIYA[1], KETAN KOTECHA[2], AJITH ABRAHAM[3], (Senior Member, IEEE), LUBNA ABDELKAREIM GABRALLA[4], AND V. SUBRAMANIYASWAMY[1]**

[1]School of Computing, SASTRA Deemed University, Thanjavur 613401, India
[2]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune 412115, India
[3]Machine Intelligence Research Laboratories (MIR Labs), Auburn, WA 98071, USA
[4]Department of Computer Science and Information Technology, College of Applied, Princess Nourah Bint Abdul Rahman University, Riyadh 11671, Saudi Arabia

Corresponding authors: R. Elakkiya (elakkiyaceg@gmail.com) and V. Subramaniyaswamy (vsubramaniyaswamy@gmail.com)

This work was financially supported by the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R178), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

**ABSTRACT** The recent developments in deep learning techniques evolved to new heights in various domains and applications. The recognition, translation, and video generation of Sign Language (SL) still face huge challenges from the development perspective. Although numerous advancements have been made in earlier approaches, the model performance still lacks recognition accuracy and visual quality. In this paper, we introduce novel approaches for developing the complete framework for handling SL recognition, translation, and production tasks in real-time cases. To achieve higher recognition accuracy, we use the MediaPipe library and a hybrid Convolutional Neural Network + Bi-directional Long Short Term Memory (CNN + Bi-LSTM) model for pose details extraction and text generation. On the other hand, the production of sign gesture videos for given spoken sentences is implemented using a hybrid Neural Machine Translation (NMT) + MediaPipe + Dynamic Generative Adversarial Network (GAN) model. The proposed model addresses the various complexities present in the existing approaches and achieves above 95% classification accuracy. In addition to that, the model performance is tested in various phases of development, and the evaluation metrics show noticeable improvements in our model. The model has been experimented with using different multilingual benchmark sign corpus and produces greater results in terms of recognition accuracy and visual quality. The proposed model has secured a 38.06 average Bilingual Evaluation Understudy (BLEU) score, remarkable human evaluation scores, 3.46 average Fréchet Inception Distance to videos (FID2vid) score, 0.921 average Structural Similarity Index Measure (SSIM) values, 8.4 average Inception Score, 29.73 average Peak Signal-to-Noise Ratio (PSNR) score, 14.06 average Fréchet Inception Distance (FID) score, and an average 0.715 Temporal Consistency Metric (TCM) Score which is evidence of the proposed work.

**INDEX TERMS** Deep learning, generative adversarial networks, sign language recognition, sign language translation, video generation.

## I. INTRODUCTION

Communication is essential for all human lives to explore their requirements and interactions with other people. Based on recent studies, various researchers found an interesting and unique style of communication in sign language across different countries. The sign languages are obviously visual cues and co-ordinate the human manual and non-manual components dramatically. It greatly supports the hard-of-hearing and speech-impaired society in getting education, jobs, and societal rights. The governments of various nations amended the multiple acts to standardize the sign language to

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

benefit the hard-of-hearing and speech-impaired community. Since, the sign language performs important role in hard-of-hearing and speech-impaired communication, the understanding and responding by the normal people requires additional training and knowledge. This creates a communication gap between ordinary people and the impaired community. The recent advancements in deep learning techniques handle such task efficiently by encompassing numerous mechanisms and mathematical approaches. The development of such systems incurs huge complexities in various phases of development, such as misclassification, self-occlusion, movement epenthesis, ambiguity, noise, and blurred output. We investigated all these challenges in a novel way to provide a better solution and aimed to build a powerful architecture to provide greater performance.

The emergence of deep learning techniques entered all the fields to exhibit their strength towards robust model development. The deep learning techniques produces impressive results in areas such as agriculture [1], anomaly detection [2], activity recognition [3], business analysis [4], [5], crop selection [6], defect monitoring [7], DNA systems [8], earth analysis [9], fraud detection [10], genomic prediction [11], human activity recognition [12], image classification [13], job matching [14], kinematic analysis [15], location prediction [16], medical systems [17], [18], [19], [20], network traffic analysis [21], number plate recognition [22], object detection [23], predictive maintenance [24], quality control [25], robotics [26], stock prediction [27], time series data analysis [28], and text generation [29], unmanned vehicle path findings [30], vehicle monitoring [31], weather forecasting [32], x-ray imaging [33], YouTube video analysis [34], zone segmentation [35]. These developments highly motivate us to pursue research in the deep learning area. Deep learning models are highly powerful and have produced intelligible achievements in a wider range of applications. However, due to the complex structures and higher number of layers, the model training process and producing the greater accuracy performances create additional challenges during the model development. These reasons cause their applicability to produce powerful models for handling complex tasks. We propose a Hybrid Deep Neural Architecture (H-DNA) which integrates the sign language recognition, translation, and video generation tasks into a single application as shown in Figure 1. We proposed a Hybrid- Deep Neural Architecture (H-DNA) which is designed to learn the different modalities of sign gestures in a signer-independent environment. This enhances the model to understand the underlying complex relationship between the input and output. The experimental results explore the effectiveness of the proposed work in terms of recognition accuracy and visual quality. In order to achieve greater flexibility and simultaneous processing of gesture sequences, we use attention mechanisms and mathematical approaches to enhance the performance of the deep model. To justify these factors, we have shown the sample output screens and outcomes of the proposed methods in section 4. The main goal of deep neural networks is to mimic

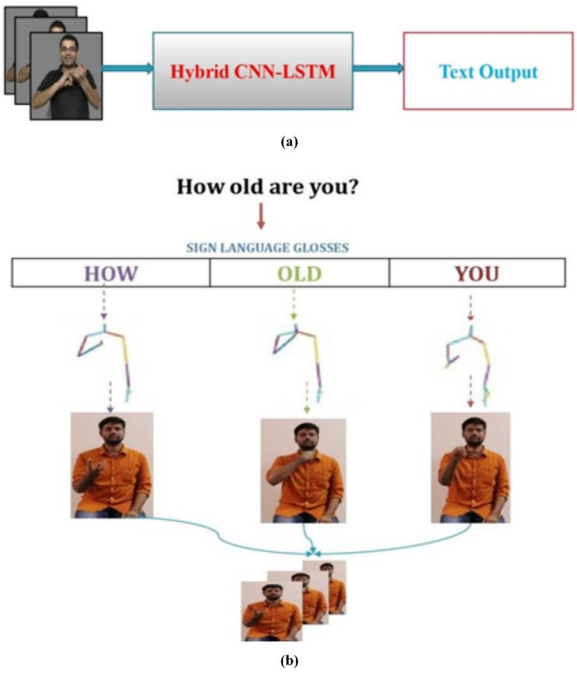


**FIGURE 1.** Overview of the proposed H-DNA framework (a) SL recognition (b) SL gesture based video generation.

the functions of the human brain to explore tremendous performance over wider tasks and diverse domain applications. The research studies on such implementations provide detailed information about the layer information, hyper parameters and advancements.

In this paper, we introduce a method to leverage the new advancements in deep neural networks to produce plausible results in translation and video generation tasks. In fact, the proposed ideology further extends to developing user interface based applications for handling real-time cases and potentially addressing the various challenges of existing approaches. Our contributions comprise the creation of Indian Sign Language (ISL) related sentence level video datasets using multiple signers without involving any specialized components like color gloves and sensors. We used a digital SLR camera and web camera devices for recording the gesture videos. The proposed H-DNA systems are capable of producing high quality videos given spoken sentence input, processing the sign gestures and translating them into spoken text. This two way mechanism is found to be superior to existing developments and comparably produces greater results in recognition and translation tasks. The experimental results have been plotted to showcase the performance of the proposed model for handling different sign corpus. We enlisted 50 student and staff volunteers to evaluate the model's performance and tabulated the scores of their evaluation by considering different parameters. Overall, the proposed H-DNA systems are designed and implemented to handle the various nuances of traditional approaches and yield better results in Sign Language Recognition and Translation (SLRT) tasks.

Although the numerous developments have made for Chinese sign Language (CSL), American Sign Language (ASL) and German Sign Language (GSL), still the performance of the model lacks in continuous cases and fails to handle the real time inputs.

The proposed H-DNA systems facilitate the real-time and accurate recognition of multimodal and multilingual sign gestures. It allows an opportunity for developing the robust applications to handle various countries based sign languages and provides solutions for communication gap exists between normal and impaired community. The proposed work has been developed as a User Interface (UI) application for handling multilingual inputs, recognizing the multimodal sign gestures, generating the sign videos, and providing accurate results over the translation and recognition tasks. To achieve the expectations, the model development underwent various stages of development to handle multimodal features and variations of multilingual sign corpuses. The proposed model has been trained using 40K videos for continuous recognition and 35K images for world level recognition tasks. The proposed system explores solutions for the real time interactions of hard-of-hearing and speech-impaired people with normal people.

The detailed investigation and various refinements of Convolutional Neural Network (CNN), Long Short Term Memory, Gated Recurrent Unit and Generative Adversarial Networks (GAN) models yield better translation results and generate high quality photorealistic videos. Sign Language plays vital role in the communication of hard-of-hearing and speech-impaired community due to their inability towards reading and writing the native language. Since the various studies dealing with SLRT research, the earlier developments have their own limitations and are still unable to be used for continuous cases. Some of the research has been known to be successful for recognizing sign language, but it requires an expensive set up and sensor devices to handle it. The tracking and recognition of specialized multimodal gesture signs is very crucial, especially in recognizing signs of different languages (multilingual). The research study on SL recognition focuses on the translation of sign gestures into English sentences and produces the text transcription for the sequence of signs. This is due to the misconception that deaf people are comfortable with reading spoken language and therefore do not require translation into sign language. To facilitate easy and clear communication between the hearing and the impaired community, it is vital to build robust systems that can translate spoken languages into sign languages and vice versa. This two way process can be facilitated using sign language recognition, translation, and video generation. With this motivation, the proposed approach intends to develop and build a novel H-DNA framework for SL recognition and translation systems as well as enhance the interactive communication between the normal and impaired community.

To the best of our knowledge, the proposed H-DNA is the first novel unified deep learning framework which addresses two different problem dimensions in SL: Sign Language Recognition (SLR) and Sign Language Translation (SLT). Using Neural Machine Translation (NMT), MediaPipe library and Dynamic GAN the proposed H-DNA will be developed for generating the high resolution videos. The proposed work simplifies the translation of spoken text to subunit signs and then defines the mapping between glosses and sign gesture images using the open pose library. Further the SL videos are produced using and DynamicGAN model. On the other hand, using CNN, LSTM and MediaPipe library, the proposed H-DNA recognizes the multilingual datasets which comprises of isolated signs and continuous sign sentences by considering multimodal features. The H-DNA was developed and implemented on GPU-powered workstations. The collection of benchmark datasets and the recording of own datasets are carried out as the first steps in implementation. To evaluate the performance of proposed H-DNA, the experimentation is performed to have three folds: The first fold deals with SL recognition, and the second fold focuses on SL video generation. The SL recognition model achieves an accuracy of not less than 98% and shows the improved performance of the proposed H-DNA. Criteria like robustness, flexibility, and scalability are considered in the third fold. We summarize the overall objectives of the proposed work as follows:

- To create & integrate heterogeneous data sources and to build a novel knowledge base consisting of multilingual and multimodal sign sentences with minimal sign glosses and skeletal level annotations by breaking down the signs into dedicated subunits.
- To augment and generate sign videos based on subunits from spoken language sentences to facilitate communication between normal and impaired (hard-of-hearing and speech-impaired) communities.
- To track and recognize the signs consisting of isolated words and continuous sign sentences including manual (one-handed and two-handed signs) and non-manual gestures in real-time scenarios.
- To build a novel application with end-to-end video generation and recognition capabilities by sharing the qualitative and quantitative results of generated sign sequences without using animated avatars or sensors, and to ensure accuracy with minimal cost.

The further discussions about the proposed model are discussed as follows. Section 2 investigates the earlier developments and provides the research gap in SLRT research and seeks the advancements in various phases of development. The proposed system details are wisely explained in Section 3 and provide sufficient details about the model development. The experimental outcome of the proposed model is shown in section 4, and finally, the conclusion and future work part summarizes the entire information about the proposed work.

## II. RELATED WORK

Sign language communication explores the powerfulness of human intelligence through hand actions and movements.

**TABLE 1.** Comparison of existing SL recognition frameworks.

| Author | Technique | Static/ Dynamic | Category | Sign Language | Isolated/ Continuous | Accuracy |
|---|---|---|---|---|---|---|
| Barbhuiya et al. 2021 [36] | CNN+SVM | Static | Alphabets and Numbers | ASL | Isolated | 99.82% |
| Aly et al. 2020 [37] | DeepLabv3+Bi-LSTM | Static | words | Arabic SL | Isolated | 89.59% |
| Lee et al. 2020 [38] | LSTM+KNN | Static and Dynamic | Alphabets | ASL | Isolated | 99.44% |
| Xiao et al. 2020 [39] | CNN+Bi-LSTM with attention | Static | Words, Sentences | CSL,GSL | Continuous | 81.22% (CSL) 76.12% (GSL) |
| Elakkiya et al. 2021 [40] | GAN+LSTM+3DCNN | Dynamic | Sentences | GSL, ASL | Continuous | 98.33% |

Despite relying on a single component (hand), it involves numerous human upper body components such as head, mouth, and gaze movements to provide a real understanding of gesture sequences in real time. Sign languages are made up of visual actions and do not have a unique pattern to identify their motion sequences. It greatly follows different styles based on its own country's nature and culture. Understanding and processing such inputs is extremely difficult for traditional machine learning approaches. It mainly supports the hard-of-hearing and speech-impaired society by getting those benefits such as education, employment, and engaging them in societal activities. There have been numerous research efforts made to produce better translation models. The real time recognition and translation of sign languages requires careful investigation of various features to produce plausible output without any misclassification and wrong sign output.

The progress Deep Learning approaches steps towards newer heights and produces fabulous results in computer vision and human action recognition applications. The introduction of hybrid models and ensembling techniques advances the capabilities of such models to handle tedious tasks. The recent research works in CNN, LSTM, GRU and GAN techniques has been investigated related to the SL recognition, translation and video generation tasks and helps to introduce the novel contributions to build a powerful framework.

The author, Barbhuiya *et al.* [36] proposed CNN+SVM based hand gesture recognition methods for static signs. This approach mainly deals with alphabets and numerals. The authors, Aly *et al.* [37] proposed a system for handling the words of Arabic SL using DeepLabv3+ gesture segmentation techniques and Bi-LSTM. The ASL recognition system for 26 alphabet level sign gesture recognition tasks is proposed by the author Lee *et al.* [38] uses LSTM with KNN techniques to provide higher recognition results. This work deals with world-level sign language communications. In addition to that, the researchers Xiao *et al.* [39] introduced continuous SL recognition using NMT approaches. The author, Elakkiya *et al.* [40] proposed an SL recognition framework using GAN+3D-CNN+LSTM Techniques. This approach utilizes the deep reinforcement learning based evaluation strategy to produce highly accurate results. The various details of the earlier literature are shown in Table 1 for exploring the new advancements with different sign languages such as American Sign Language (ASL), Chinese Sign Language (CSL) and German Sign Language (GSL). The conventional sensor-based approaches demand extra equipment to be worn by the signer. The use of data gloves, color gloves, depth cameras, and leap motion controllers creates additional overhead for the signer to communicate normally and poses huge limitations [41]. Although it gives good prediction results, drastically loses the scope in real time applications. In addition to that, it creates discomfort for the child and normal people during the conversation.

The optimization of hyper parameter values and the imposing of various constraints produces plausible outcomes and attracts the researchers. The primary version of the CNN model is introduced by authors Chen and Koltun [43] produces images from semantic layouts. The model investigates the different loss functions and produces photographic results. The model performance bottlenecks while handling the large scale of images and adds the various intrinsic challenges. Similarly, the researchers in Oord et. al. [44] discussed the development of gating mechanism based PixelCNN models for image generation. The model has been evaluated using the datasets CIFAR-10 and ImageNet. Since the model applies different conditions on embedding features to produce quality image generation results, and extending the performance for videos creates additional overheads. Although numerous advancements were made in the research work [45], the production of sign gesture videos is blurred and spatial details are incoherent.

The development of ambient models such as FUNIT [46], StarGAN [47], StarGAN v2 [48], MoCoGAN [49], LPGAN

[50], InfoGAN [51], pix2pix [52], and CycleGAN [53] deals with the image generation and video production tasks efficiently. Since SL communication involves the various manual and non-manual cues of humans and their facial, eye, gaze, and mouth expressions, it demands some advancement in the earlier approaches. In addition to that, the ordering of gesture sequences greatly varies from the English sentence order. In order to address these aforementioned challenges, we introduce a novel approach for aligning the frame sequences and generating the intermediary frames between the sign gesture images. The proposed model deals with the various nuances of SL gestures and its components and produces plausible outcomes. The GAN networks are found to be highly capable of producing plausible results across a wider range of diverse domain applications. The applications such as security [54], [55], baggage inspection [56], infected leaf identification [57], covid-19 prediction [58], agriculture [59], business process monitoring [60], Brain MRI synthesis [61], flood visualization [62], estimating the standards of gold [63], ECG wave synthesis [64], Internet of Things (IoT) [65] and Dengue Fever sampling [66].

The two major components of GAN networks are generator and discriminator, and they play a vital role in image or video generation. The discriminating capability of the discriminator helps to produce high quality videos in diverse domains and is further investigated in the proposed work for qualitative production of sign gesture videos. The incorporation of CNN models with conditional GAN networks produces drastic improvements in video generation quality and efficiently handles the various traits of details present in an image or video. Based on the discriminator classification, the generator networks underwent the fine-tuned training process to produce photorealistic results. The authors, Mirza and Osindero [67], introduced the conditional-based GAN network model by applying constraints on label information. We use this approach in our work to produce videos based on the conditioned labeling approach. The advent of DynamicGAN models addresses the existing challenges by using strided mechanisms in convolution operations to produce improved results. The video generation process using GAN networks encompasses the additional approaches to produce photo-realistic videos and keeps the coherent spatial details clear.

Although there are enormous research going on in the field of Sign Language Recognition, translation and generation systems the existing systems still face a lots of challenges. The primary challenge with the Sign Language recognition and generation system is the lack of availability of large-scale open-source Indian Sign Language Dataset with natural conditions. To overcome this issue we have developed a multi-signer, multi-model Sign Language dataset and have provided it as open-source resource for further research purposes. For Sign Language recognition systems, we have built the recognition model in such a way that it detects the signs irrespective of the complex backgrounds, multi-modality, signer skin tone, signer clothing constraints, sign speed etc., which are the major drawbacks in the existing systems. In case of Sign Generation, we have considered the limitation to small size vocabulary, model performance improvement, low model complexity, proper alignment of the key points, signs in spatial domain etc.

## III. THE PROPOSED H-DNA SYSTEM

The proposed hybrid H-DNA framework model comprises various phases of development, such as SL recognition, multilingual sentences into sign word conversion, pose estimation using MediaPipe, and SL video generation. The proposed H-DNA framework aims to integrate all these modules and provide a real time solution to SLRT research challenges. Neural Machine Translation (NMT) is the process of translating sentences from one language into another. It uses artificial neural networks to yield highly translatable results. The identification of human poses in images or videos is performed using the mediapipe library. It helps to predict the various poses of humans in various environments. Pose estimation is based on a number of key points on the human body.It uses the Parity Affinity Fields approach to implement it. The VGG-19 model is used for classifying the different gesture styles. It uses different $3 \times 3$ filters in the convolution layers. The convolution layers provide a feature map by scanning the image features. The role of pooling layers is to reduce the information generated by the convolution layers. To vectorize the output as a single array, the fully connected layer is used. The incorporation of dynamic GAN [86] provides high quality video generation results by encompassing the various approaches such as frame generation and video completion techniques. The LSTM network is used for predicting the text equivalents of the sign gestures and further helps to produce the language sentences. The following subsections explore the various technical details and summarize the powerfulness of each technique.

This section explains the implementation details of the proposed H- DNA framework. In the first fold, we developed the SL recognition model using the MediaPipe library and the VGG-19 model. Furthermore, we incorporate the Bi-LSTM network for text generation. In SLR, the input of continuous gesture sequences is processed by the MediaPipe library to capture the pose sequences, angle between fingers, hand movements and locations, orientations, mouth expressions, and facial actions. Based on these key points, the VGG-19 model estimates the class of gestures. The incorporation of CNN and LSTM networks in such a hybrid way produces higher recognition accuracy and noticeable performance. The temporal details are analyzed sequentially to predict the translation text without any misclassification.

We trained our model using 40,000 videos for 320 classes to provide wider support over multilingual sign corpus comprises of multimodal features. The sample gesture images of our own created ISL-CSLTR dataset [42] are shown in Figure 2 and greatly support the ISL-related SLRT research. In general, the SL video generation process is treated as a highly intensive task due to the production of sign gesture


**FIGURE 2.** Sample word level sign gesture images of ISL-CSLTR dataset.

videos from English sentences. The qualitative production of SL videos for the new input sentences poses various levels of difficulties by considering the manual and non-manual cues of the signers. Such a translation process demands more attention at each step to produce high quality results. The emergence of various deep generative models has advanced and secured new milestones in photorealistic image generation and video production.

### A. SL RECOGNITION

In the first phase, the development of SL recognition using hybrid CNN+Bi-LSTM techniques is carried out. The main objective of this hybrid approach is to sequence predict in SL videos. The CNN layers are used for gesture class identification and LSTM networks for predicting the class sequence. The combination of these two networks processes the spatio-temporal details of SL input videos and produces the text output. The first segment uses CNN layers and is further utilized by the Bi-LSTM networks with dense layers to yield plausible results. We used the VGG-19 model [68], which consists of 16 convolution layers and 3 fully connected layers. The CNN network processes images of size 254 $\times$ 254 and the first and second layers are convolutional layers. It uses 3 $\times$ 3 filters with stride level 1. The max pooling operation is performed using stride level 2 and a window size of 2 $\times$ 2. After this process, the dimensions of pixels are reduced to 112 $\times$ 112 $\times$ 64. Further, the convolution layer of varying filter size 128, 56, 28 is applied and reduces the size of the image as well as focuses the important features. The fully connected layer summarizes all classes of inputs and produces the probability of prediction values using the softmax layer. The network is trained to handle 35K images of 192 classes representing different gesture poses based images for different words. After completion of preprocessing steps, the videos of high resolution to be 1920 $\times$ 1080 and converted into numpy arrays for easier processing using skvideo packages. Each class of sign gestures is recorded with 50 repetitions to provide better learning and prediction performance of the model. The key points based pose information is captured parallel to maintain the gesticulation details and aids the better assistance over classification and prediction tasks. The incorporation of the CNN based pre-trained model VGG-19 helps to automate the SL recognition tasks in a better way. The various basic operations, such as convolution operations and max pooling, are applied repeatedly to learn the finer details of the images. The VGG-19 model provides better classification results than the multilingual sign language datasets. The results are passed to the Bi-LSTM networks to predict the target sentences matches with the video sequences. The intermediary feature map results of the proposed hybrid CNN+Bi-LSTM model are shown in Figure 3.

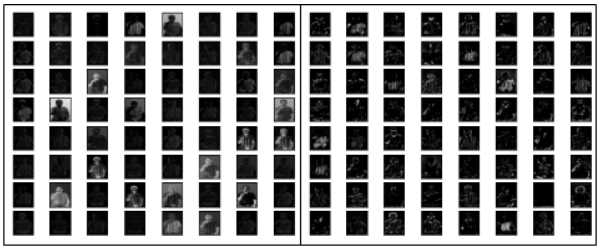

**FIGURE 3.** Visualization of the intermediate feature map results of VGG-19 network.

The VGG-19 model produces the vector representation of images and classification results. Based on such input, the LSTM layers process the information and generate the textual descriptions. In this context, the textual descriptions are language sentences that match with gesture sequences. The entire CNN model is handled by the time distribution layers to handle multiple inputs for different time steps. The LSTM units apply back propagation to tune the hyper parameters such as learning rate, batch sizes. The weight and bias values are also updated to build a powerful framework. We set the learning rate value as 0.01 and the batch size as 64. The LSTM networks [69] are found to be powerful components in text generation, image captioning, and machine translation tasks. The LSTM network has three gates: (i) input gate, (ii) output gate, and (iii) forget gate. The separate memory cell is added and handles a higher number of layers than GRU. The forget gate decides the kind of information to be discarded from memory and uses sigmoid activation functions to squish the values between zero and one. Due to this functionality, the values multiplied by 0 become zero and can be easily removed. The input gate updates the cell state for processing the new inputs. The memory cells remain the amount of information for time stamp t. The output gate finalizes the information to be output from the model. The forget gate functions are represented using the following Equation (1) The cell state is a key for the LSTM network, passing through the entire chain link of LSTM modules and governed by the aforementioned three layers. The forget gate decides the information to be thrown away from the memory. The role of input layers is to provide the desired inputs and update the cell state values. The output layers produce the text results. We use sigmoid and tanh activation functions to

produce plausible outcomes. We use a bi-directional LSTM approach to focus on the text generation tasks efficiently. The proposed hybrid CNN + Bi-LSTM techniques based SL recognition system architecture is shown in Figure 4.

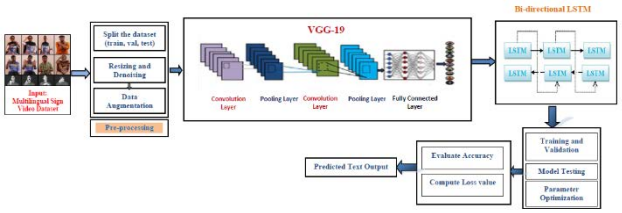


**FIGURE 4.** SL recognition system.

The LSTM network is envisioned as a strong method to handle sequential tasks. It provides a solution to the vanishing gradient problem. Since it handles the longer sequential inputs, which are applied in domains such as image captioning, text generation and time series based applications. The LSTM network was introduced by the authors, Cho *et al.* [70]. The forget gate ($fr_t$) operations are represented using Equation 1. It decides the information to be discarded from cell states by applying the sigmoid activation function. The value 1 represents keeping the information and 0 denotes its removal. The general equations describing the various operations of the LSTM Network are stated in Equation 1.

$$fr_t = \sigma\left(W_{fr} \cdot [h_{t-1}, In_t] + b_{fr}\right) \tag{1}$$

The next step processes the sequence of inputs and decides the next information to be fed into the cell state. The input layer represented using Equation 2 denotes the next value to be updated in the cell state. Next, Equation 3 represents the vector values of candidate results.

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, In_t] + b_i\right) \tag{2}$$

$$Ca_t' = \tanh\left(W_{Ca} \cdot [h_{t-1}, In_t] + b_{Ca}\right) \tag{3}$$

The update of new values ($\mathbf{Ca_t}$) by using multiplication operations and the refinement of old cell state values takes place using Equation 4.

$$Ca_t = fr_t * Ca_{t-1} + i_t * Ca_t' \tag{4}$$

The output gate operations are denoted using Equation 5 and Equation 6. It decides the information to be passed as output.

$$out_t = \sigma\left(W_{out} \cdot [h_{t-1}, In_t] + b_{out}\right) \tag{5}$$

$$h_t = out_t * \tanh(Ca_t) \tag{6}$$

The proposed hybrid CNN + Bi-LSTM Technique is shown in Figure 5. The detailed steps of our implementation and provides the step-by-step procedures. It gives a detailed overview of the execution of VGG-19 model training. The LSTM network operations to produce the language sentence output are clearly elaborated in the rest of the sections.

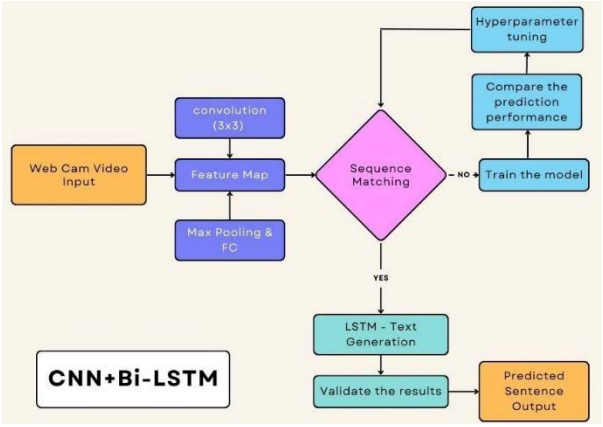


**FIGURE 5.** Flow chart for hybrid CNN+Bi-LSTM technique.

The LSTM network utilizes the memory cell component explicitly, and the cell states regulate the kind of information to be kept or discarded from memory. During each iteration cycle process, the LSTM network processes the previous hidden state values (ht-1), current input values (Int) and the previous cell state values (ht). The parameters weight and bias vectors are updated regularly during the back propagation process to produce accurate translation results. We use Adam optimizer and drop out regularization techniques to obtain greater results over the benchmark datasets.

### *B. MULTILINGUAL SENTENCES INTO SIGN WORD CONVERSION*

This section explains the translation process of language sentences into sign words using the NMT and attention mechanism. The conventional NMT techniques have proven to have appreciable performance in language translation tasks. We use a hybrid NMT + Attention mechanism for translating the multilingual sentences into sign words. The NMT technique uses RNN and its variants to process the longer sequences and produces better results in different domain applications. We introduce the novel deep-stacked GRU technique in machine translation tasks to achieve greater translation results over multilingual input sentences. The translation process is carried out using the following steps: The first step deals with the text preprocessing of the spoken sentences. The spoken sentences are cleaned by removing the special characters, punctuation marks, and symbols. We add the <START> token at the beginning of the sentences and <END> tokens at the end of the sentences. This approach benefits the model learning process of where to start and stop. The word embedding techniques are used to convert the tokens into dense vectors and pass them to the next level. The proposed deep-stacked GRU technique efficiently handles the translation tasks and produces accurate results. The GRU networks use two gates: (i) the update gate and (ii) the forget gate. The update gate governs the information to be newly added and the forget gate regulates the information to be kept

or thrown away. The following equations clearly explore the various operations of GRU units.

The deep-stacked GRU units are chain-link based on different modules which are executed iteratively in order to produce the sequential outputs. The input value from the current step is denoted as $x_t$ and the input of previous hidden layers is represented as $h_{t-1}$. The operations of the update gate ($Z_t$) are represented using Equation 7.

$$Z_t = \sigma\left(W^{(z)}x_t + U^r h_{t-1}\right) \tag{7}$$

The current input value ($x_t$) and the weight (W) values are multiplied in the first part, and the second part multiplies the previous hidden state values ($h_{t-1}$) and its weights (U) and finally the values are summed up to provide the new values to the update gate. The sigmoid ($\sigma$) activation function is applied over the resultant values to round up the prediction results in the range of zero to one. The update gate concludes the volume of information to be passed to the next state. The reset gate decides the removal of information based on the importance of particular vector towards the prediction of next sequences. The executions of reset gate are demonstrated using the Equation 8 as follows.

$$r_t = \sigma\left(W^{(r)}x_t + U^r h_{t-1}\right) \tag{8}$$

The reset gate ($r_t$) combines the results of the multiplication operation performed on the input ($x_t$) and weight (W) values as well as the previous hidden node values ($h_{t-1}$) and its weight values (U). The sigmoid activation is applied to the results. The current values ($h_{cur}$) to be present in the memory unit are computed using Equation 9.

$$h_{cur} = \tanh\left(Wx_t + r_t \odot Uh_{t-1}\right) \tag{9}$$

The current and previous node values are multiplied with weight values. The Hadamard product, known as element-wise multiplication, is performed over the reset gate and previous hidden states values. Finally, the non-linear activation function tanh is computed on the final outcome. The last step results in being recorded in memory units ($h_f$) at time step t is computed using Equation 10.

$$h_f = Z_t \odot h_{t-1} + (1 - Z_t) \odot h_t^{'} \tag{10}$$

The deep stacked approach provides better results over a wider range of applications and reduces the computational complexity of the model drastically. The deep stacked GRU has several units of GRU blocks and performs the model training in parallel. The detailed structure of deep stacked GRU units is depicted in Figure 6.

Further, we incorporate the attention mechanism proposed by Bahdanau *et al.* [71]. The attention mechanism focuses on the particular context in encoder unit matching with target translation to yield high quality results. The cyclic execution of the deep stacked GRU units is shown clearly in Figure 7.

The GRU units process the spoken sentences input using encoder and decoder based approach. The encoder network of GRU processes the source format of input sentences.

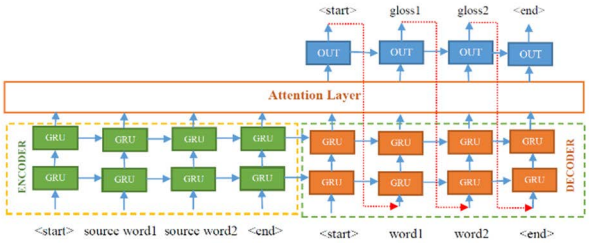


**FIGURE 6.** Proposed deep stacked GRU system.

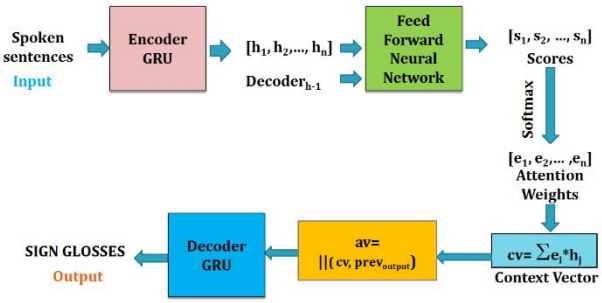


**FIGURE 7.** Execution flow of GRU based encoder decoder system.

The dense vector values of each word are passed to a feed forward neural network to learn the source representation. The proposed hybrid NMT model handles varying lengths of sentences and changes the translation results accordingly. The score values produced by the networks are further processed by the softmax function and yield attention weights. The context vectors are calculated by multiplying the attention weight values and hidden state values.

We incorporated the attention mechanism proposed by the researcher Bahdanau *et al.* [71] to yield the accurate translation results. The attention vector is estimated by concatenating the context vectors and previous output. Finally, the decoder network produces the target sign gloss output. The proposed hybrid NMT + Attention model is evaluated using the three benchmark sign corpus datasets such as RWTH PHOENIX Weather 2014T dataset [72], How2Sign Dataset [73], and ISL-CSLTR Dataset [41] and the results are shown in section 4. The computation of attention weights is done using Equation 11.

$$\alpha_{ts} = \frac{\exp(score(h_t, \bar{h}_s))}{\sum_{s'=1}^{S} \exp(score(h_t, \bar{h}_{s'}))} \tag{11}$$

The context vector is calculated by using Equation 12.

$$c_t = \sum_s \alpha_{ts}\bar{h}_s \tag{12}$$

The Bahdanau's attention vector is calculated by using Equation 13.

$$a_t = f(c_t, h_t) = \tanh\left(W_c\left[c_t; h_t\right]\right) \tag{13}$$

The proposed Deep stacked GRU algorithm uses stacked layers of GRU to effectively process the sequential inputs and

translate them into target form. We apply the Bahdanau *et al.* [71] attention mechanism to compute distinct context vector values and get good results. The recursive nature of GRU processes the entire source sentences and translates them into target sentences. We use beam size 10 and tanh and sigmoid activation functions. The proposed model totally processes 40k sentences by combining multilingual sign corpus collected from different sources.

### C. POSE ESTIMATION USING MEDIAPIPE

The MediaPipe library was developed to provide human pose estimation results over image and video files. This framework is stated as an impressive one to track the details of human activity in public environments, sign gesture pose recognition, fraud monitoring, and yoga pose analysis. We use the MediaPipe library to estimate the poses of different signers and key points, which are used for generating the new poses using the deep generative networks. The sample results of the MediaPipe library are shown in Figure 8.

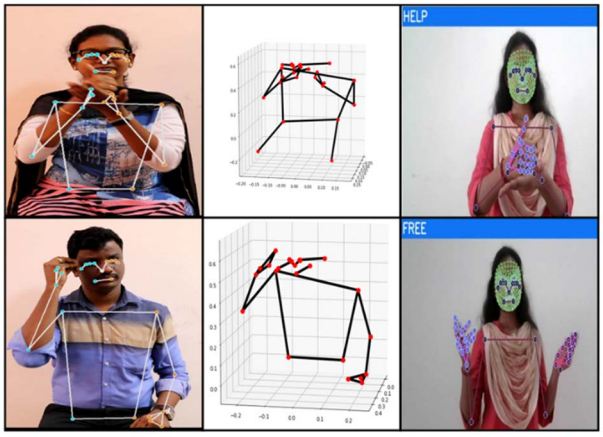


**FIGURE 8.** Sample human pose estimation results and 3D plots using MediaPipe library.

### D. SL VIDEO GENERATION

The sign gesture video generation tasks are performed using deep generative models. We introduce the novel DynamicGAN network for producing plausible, photo-realistic high quality videos. The video generation involves a series of step-by-step approaches to produce high-quality results. We carefully investigated the various mathematical models and deep generative frameworks to develop the novel framework. The advancements of GAN networks have found them proficient in generating high quality images and videos. The GAN networks synthesis the medical images efficiently as well. We incorporate the conditional GAN model [67] as the basic framework for our proposed DynamicGAN model. Furthermore, we use the VGG-19 pre-trained CNN network for sign gesture classification. The techniques such as intermediary frame generation, deblurring and image alignment, pixel normalization, video completion are added additionally with the generator network to produce the photo-realistic high quality sign gesture videos. The integrated architecture for translating the multilingual sentences to sign video generation is shown in Figure 9.

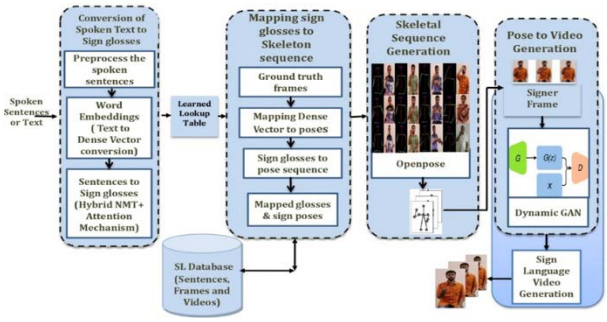


**FIGURE 9.** SL gesture video generation using NMT+openpose+GAN techniques.

The GAN network consists of two units known as the generator unit and the discriminator unit. The generator unit produces the new images or videos from the noise distribution of real data. The latent space provides various details of real data based on which, it produces the new images or videos. The conditional GAN model uses conditioned labels to produce the sharp images. The generated results are verified by the other unit known as the discriminator. The discriminator unit classifies the real and fake samples as shown in Figure 10.

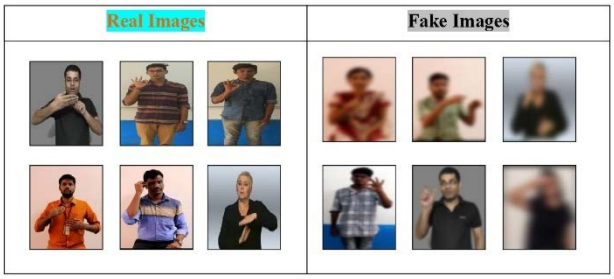


**FIGURE 10.** Discriminator network classification results.

Depending on the classification results, the generator networks fine tune their performance to produce plausible images and videos. We use a U-Net-like framework [74] for learning the structure of real data distribution. From which, the target pose images are generated quantitatively. The encoder network performs the convolution function, batch normalization and activation function for Leaky ReLU. The decoder network utilizes the transposed convolution function, batch normalization techniques, dropout regularizer, and finally, ReLU activation functions. The loss value for the generator network is computed using the sigmoid cross-entropy loss. Further, the L1 loss calculates the mean absolute error between the real and generated results and aids in producing high quality results. The Discriminator unit incorporates the PatchGAN [52] classification techniques to discriminate the real and fake samples. The Convolution

function, Batch normalization, and the Leaky ReLU activation function are applied sequentially to produce plausible outcomes. The discriminator network estimates the realness of the generated results. It uses sigmoid cross-entropy loss function to measure the quality of generated results compared with real ones. The proposed Dynamic GAN model is implemented in high end GPU based environment. The Dell Precision 7820 Tower workstation is used to accomplish the entire development process. It comprises pairs of Intel Xeon Silver 4210 2.2. GHz processors and 10 cores. The Nvidia Quadro RTX40000 provides GPU support for model training. We use batch normalization and Adam optimization techniques with the values $\alpha = 2\text{e-}4$, $\beta = 0.5$ and $\beta = 0.999$. We set the batch size value as 128, dropout is 0.01 and initial learning rate as 0.01. The Leaky ReLu value is set as 0.1 and ReLu activation functions are further applied. The mini batch size is set as 100 and the momentum is 0.05. The proposed DynamicGAN framework is experimented using the multilingual sign corpus such as RWTH-PHOENIX-Weather 2014T dataset, ISL-CSLTR dataset, and How2Sign dataset. The results are shown in section 4.

We use the Mean Squared Error (MSE) Metric to evaluate the loss values in the generator network outcomes stated in Equation 14.

$$\mathcal{L}_{\text{MSE}}(\text{xt}, \text{gn}) = \ell_{\text{MSE}}(\text{G}(\text{X}_{\text{xt}}), \text{gn}) = \|\text{G}(\text{xt}) - \text{gn}\|^2 \tag{14}$$

The Sigmoid Cross-Entropy loss combines the activation function sigmoid as well as the Cross-Entropy loss function. Due to the independent execution of these loss functions, it does not affect the results of one on another.

### E. DATASET

RWTH-PHOENIX-Weather 2014T dataset: This dataset deals with the SLRT research for German sign language [72]. It consists of 40k videos for sentence level. The videos are recorded using 9 native signers.

ISL-CSLTR dataset: The ISL-CSLTR dataset was published by the researchers [41] to conduct the SLRT research in Indian sign language. It consists of 700 videos for 100 sentences each. The videos are recorded using seven different signers. How2Sign dataset: The How2Sign dataset [73] contains the SLRT research for American Sign Language. It consists of 2,456 videos for sentences. The videos were recorded using 11 different signers.

## IV. EXPERIMENTAL RESULTS

This section provides the experimental results of various phases of development, which are performed and investigated to build a complete framework for SLRT research challenges. The proposed H-DNA framework functionalities are tested in different stages of the development cycle. In addition to that, we have shown the user interface screens of the final application. During the first phase of development, the SL recognition model is implemented using hybrid CNN+Bi-LSTM techniques. The proposed model has been trained and validated to produce better results. We inputted 25k images for training and 5k images for validation purposes. The model performance is shown in Figure 11. The proposed model achieves significant improvements in classification accuracy and recognition performance. Furthermore, we plot the confusion matrix for obtaining the classification performance. The confusion matrix results are shown in Figure 12. This demonstrates the improved performance of the proposed hybrid CNN-LSTM model.

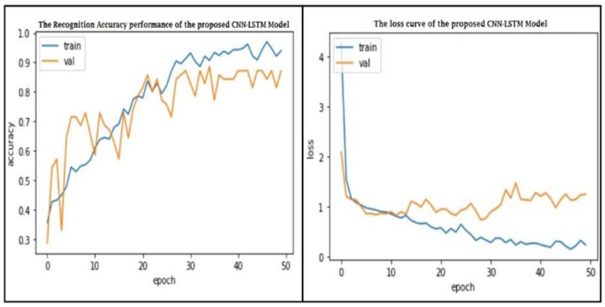


**FIGURE 11.** Accuracy and loss evaluation of hybrid CNN-LSTM model.

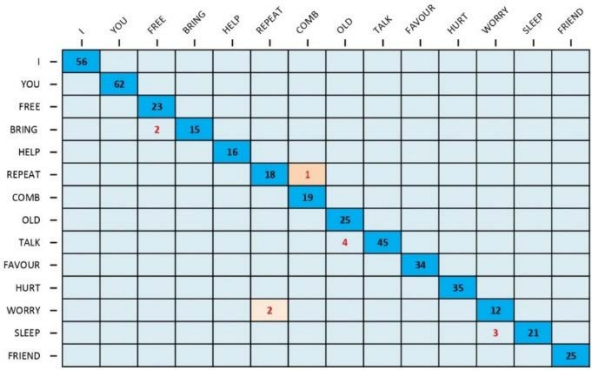


**FIGURE 12.** Confusion matrix results of the hybrid CNN-LSTM Model.

The classification performance of the proposed hybrid CNN-LSTM model is evaluated using the following metrics. The accuracy of the proposed model is compared with the existing work and the comparison results are tabulated in Table 2.

We further investigated the proposed hybrid CNN+Bi-LSTM model performance using the following equations. The precision is computed using the Equation 15, the Recall is calculated using Equation 16, F1 Score is calculated using the Equation 17 and the accuracy is computed using the Equation 18 where TP, TN, FP, FN denotes true positive, true negative, false positive and false negative values. The various quality metrics are computed using the following Equations and the results are tabulated in Table 3.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

**TABLE 2.** Comparison of existing SL recognition models with hybrid CNN-LSTM model.

| Author | Method | Accuracy (%) |
|---|---|---|
| Hao et al. 2021 [75] | SMKD | 71.3 |
| Dias et al. 2021[76] | Random Forest | 74.5 |
| Rastgoo et al. 2020 [77] | SSD+2D-CNN+3D-CNN+LSTM | 76.7 |
| Li et al. 2020 [78] | I3D | 79.1 |
| Camgoz et al. 2020 [71] | Transformer + CTC | 79.3 |
| Barbhuiya et al. 2021[36] | CNN+SVM | 80.1 |
| Aly et al. 2020 [37] | DeepLabv3+Bi-LSTM | 82.2 |
| Lee et al. 2020 [38] | LSTM+KNN | 84.4 |
| Xiao et al. 2020 [39] | CNN + Bi-LSTM with attention | 93.4 |
| Natarajan et al. 2022 (a) [85] | NMT+GAN | 96.3 |
| Natarajan et al. 2022 (b) [86] | Openpose+GAN | 97.1 |
| **Hybrid VGG-19+ LSTM** | | **98.5** |

$$\text{F1 Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{17}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{18}$$

The performance of the hybrid NMT + Attention model is evaluated using the BLEU metrics depicted in Figure 13. It shows the performance of the proposed hybrid NMT + Attention model compared with existing work. Further, the performance of the hybrid NMT + Attention model is analyzed using the attention plots depicted in Figure 14. The attention plot shows the real translation performance of the model by comparing the source and target sentences. The blocks are highlighted in white color representing the role of attention mechanism in the context of particular word translation.

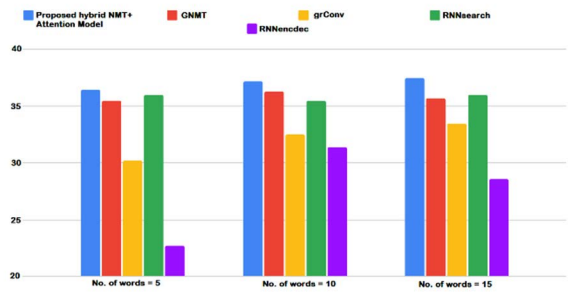


**FIGURE 13.** Hybrid NMT+Attention model evaluation results using BLEU score metric by considering the various word length.

**TABLE 3.** Hybrid CNN-LSTM model performance comparison using existing frameworks.

| Framework | Precision value | Recall value | F1-Score | Accuracy |
|---|---|---|---|---|
| SMKD [75] | 0.9664 | 0.8312 | 0.8521 | 0.9233 |
| Random Forest [76] | 0.8426 | 0.9432 | 0.9735 | 0.9217 |
| SSD+2D-CNN+3D-CNN+LSTM [77] | 0.9126 | 0.9352 | 0.8714 | 0.9445 |
| I3D [78] | 0.9451 | 0.8174 | 0.9254 | 0.9112 |
| Transformer + CTC [71] | 0.9634 | 0.8252 | 0.8735 | 0.9314 |
| CNN+SVM [36] | 0.9335 | 0.9453 | 0.9125 | 0.9455 |
| DeepLabv3+Bi-LSTM [37] | 0.8746 | 0.9454 | 0.9524 | 0.8552 |
| LSTM+KNN [38] | 0.9558 | 0.9646 | 0.9586 | 0.8523 |
| CNN + Bi-LSTM with attention [39] | 0.9652 | 0.9526 | 0.9642 | 0.9512 |
| **Hybrid VGG-19 + Bi- LSTM** | **1.0000** | **0.9843** | **0.9758** | **0.9856** |

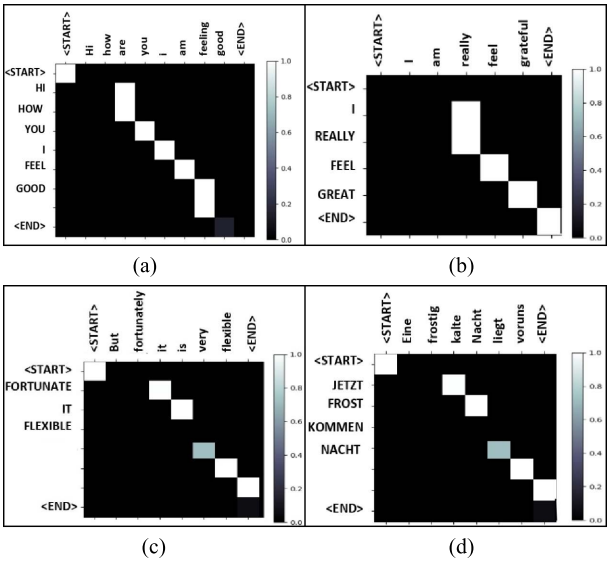


**FIGURE 14.** (a), (b) Attention plot results for ISL-CSLTR dataset, (c) How2Sign dataset, (d) RWTH-PHOENIX-Weather 2014T dataset.

We compared the proposed Dynamic GAN model performance in terms of quality and quantity by experimenting with multilingual sign language datasets and the results are shown below. Figure 15 depicts the video generation results of RWTH-PHOENIX-Weather 2014T dataset, Figure 16 shows the video generation results of the ISL-CSLTR dataset and Figure 17 depicts the video generation results of RWTH-PHOENIX-Weather 2014T dataset. We further compared the proposed Dynamic GAN model with existing deep generative

**FIGURE 15.** Video generation results of DynamicGAN model for RWTH-PHOENIX-Weather 2014T dataset.

**FIGURE 16.** Video generation results of DynamicGAN model for ISL-CSLTR dataset.

**FIGURE 17.** Video generation results of DynamicGAN model for How2Sign dataset.

models. The quantitative evaluation is carried out using the benchmark sign corpus. The results show the improved performance of our approach compared with existing models. Table 4 depicts the performance of the proposed model compared with existing models in terms of realism, relevance, and coherence using human evaluators. We validated the generated frame quality and temporal coherence using FID2Vid scores shown in Table 5.

The Structural Similarity Index Measure (SSIM) metric represented using Equation 19 is used for assessing the image quality. We use the SSIM metric for comparing the model's performance with existing approaches. This metric assesses the structural information degradation of generated video frames and the results are shown in Table 6.

$$\mathrm{l}(\mathrm{x}, \mathrm{y}) = \frac{2\mu_{\mathrm{x}}\mu_{\mathrm{y}} + \mathrm{C}_1}{\mu_{\mathrm{x}}^2 + \mu_{\mathrm{y}}^2 + \mathrm{C}_1} \tag{19}$$

The proposed DynamicGAN model performance has experimented with inception score metrics. The high score denotes

**TABLE 4.** Comparison of DynamicGAN model performance with existing works by human evaluation metrics.

| Model | Realism | Relevance | Coherence |
|---|---|---|---|
| EBGAN [79] | 3.72 | 4.72 | 4.58 |
| StackGAN [80] | 3.69 | 4.68 | 5.68 |
| BEGAN [81] | 3.86 | 4.86 | 4.92 |
| PROGAN [82] | 4.56 | 3.56 | 4.56 |
| InfoGAN [51] | 5.32 | 5.32 | 5.38 |
| SGAN [83] | 6.1 | 5.1 | 6.1 |
| BiGAN [84] | 7.72 | 6.72 | 7.72 |
| **DynamicGAN** | **7.91** | **7.46** | **8.46** |

**TABLE 5.** Performance Comparison of DynamicGAN model for FID2vid metrics.

| Model | FID2vid |
|---|---|
| EBGAN [79] | 4.28 |
| StackGAN [80] | 5.78 |
| BEGAN [81] | 4.62 |
| PROGAN [82] | 4.46 |
| InfoGAN [51] | 5.58 |
| SGAN [83] | 5.3 |
| BiGAN [84] | 4.62 |
| **DynamicGAN** | 3.46 |

**TABLE 6.** Cmparison of Structural Similarity Index Measure (SSIM) metric.

| Framework | SSIM | | |
|---|---|---|---|
| | RWTH-PHOENIX-Weather 2014T dataset | ISL-CSLTR dataset | How2Sign dataset |
| EBGAN [79] | 0.702 | 0.802 | 0.856 |
| StackGAN [80] | 0.785 | 0.810 | 0.863 |
| BEGAN [81] | 0.852 | 0.826 | 0.796 |
| PROGAN [82] | 0.891 | 0.901 | 0.892 |
| InfoGAN [51] | 0.863 | 0.865 | 0.892 |
| SGAN [83] | 0.836 | 0.796 | 0.783 |
| BiGAN [84] | 0.785 | 0.693 | 0.782 |
| **DynamicGAN** | **0.901** | **0.937** | **0.925** |

the model's performance over multiple domains and the generation capability of the generator. The computation of IS is performed using the following Equation 20.

$$\mathrm{IS}(\mathcal{G}) = \exp\left(\mathbb{E}_{\mathrm{x}\sim \mathrm{p_g}} \mathcal{D}_{\mathrm{KL}}\left(\mathrm{p}\left(\mathrm{y} \mid \mathrm{x}\right) \| \mathrm{p}(\mathrm{y})\right)\right) \tag{20}$$

Let x denotes the generated images of the generator network G, let $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$ denotes the class distribution of generated

**TABLE 7.** Comparison of inception score of dynamic GAN model.

| Framework | Inception Score | | |
|---|---|---|---|
| | RWTH-PHOENIX-Weather 2014T dataset | ISL-CSLTR dataset | How2Sign dataset |
| EBGAN [79] | 12.32 | 13.5 | 13.62 |
| StackGAN [80] | 14.13 | 15.3 | 13.63 |
| BEGAN [81] | 14.65 | 13.9 | 13.65 |
| PROGAN [82] | 12.2 | 13.1 | 14.0 |
| InfoGAN [51] | 11.68 | 12.36 | 14.6 |
| SGAN [83] | 13.1 | 12.12 | 10.53 |
| BiGAN [84] | 12.36 | 10.23 | 9.32 |
| **DynamicGAN** | **8.4** | **8.6** | **8.2** |

**TABLE 8.** PSNR Score evaluation for multilingual sign corpus.

| Framework | PSNR Score | | |
|---|---|---|---|
| | RWTH-PHOENIX-Weather 2014T dataset | ISL-CSLTR dataset | How2Sign dataset |
| EBGAN [79] | 26.32 | 28.5 | 26.62 |
| StackGAN [80] | 27.13 | 26.3 | 26.63 |
| BEGAN [81] | 25.65 | 27.9 | 27.65 |
| PROGAN [82] | 27.2 | 28.1 | 28.0 |
| InfoGAN [51] | 28.68 | 28.36 | 28.6 |
| SGAN [83] | 27.1 | 27.12 | 28.53 |
| BiGAN [84] | 28.36 | 28.23 | 27.32 |
| **DynamicGAN** | **29.4** | **30.6** | **29.2** |

samples and the marginal probability function, denoted as $\mathbf{p}(\mathbf{y})$. The Inception score results are depicted in Table 7.

The PSNR metric provides a comparison result between real and generated results. The high PSNR indicates the improved quality of the generated results. The PSNR metric is compared between different sign corpus for analyzing the proposed DynamicGAN Model performance. The results of the PSNR metric are shown in Table 8 and calculated using Equation 21 and Equation 22.

$$\mathrm{PSNR}\,(\mathrm{gt},\mathrm{ge}) = 10\log_{10}\left(\frac{255^2}{\mathrm{MSE}\,(\mathrm{gt},\mathrm{ge})}\right) \tag{21}$$

$$\mathrm{MSE}\,(\mathrm{gt},\mathrm{ge}) = \frac{1}{\mathrm{MN}}\sum_{\mathrm{i}=1}^{\mathrm{M}}\sum_{\mathrm{j}=1}^{\mathrm{N}}\left(\mathrm{gt}_{\mathrm{ij}} - \mathrm{ge}_{\mathrm{ij}}\right)^2 \tag{22}$$

The Frechet Inception Distance (FID) metric is used to assess the quality of generated video frames and is computed using Equation 23. The quality of pixels and temporal consistency are measured. The lowest FID scores indicate better results. The mean and covariance values are computed to compare

**TABLE 9.** Fréchet inception distance (FID) metric evaluation for multilingual sign corpus.

| Framework | Fréchet Inception Distance (FID) | | |
|---|---|---|---|
| | RWTH-PHOENIX-Weather 2014T dataset | ISL-CSLTR dataset | How2Sign dataset |
| EBGAN [79] | 35.42 | 14.5 | 16.62 |
| StackGAN [80] | 34.23 | 14.3 | 13.23 |
| BEGAN [81] | 35.65 | 36.9 | 39.25 |
| PROGAN [82] | 35.2 | 34.1 | 41.0 |
| InfoGAN [51] | 29.68 | 36.36 | 39.6 |
| SGAN [83] | 25.12 | 25.12 | 22.23 |
| BiGAN [84] | 18.36 | 18.23 | 13.32 |
| **DynamicGAN** | **14.4** | **15.5** | **12.3** |

**TABLE 10.** Temporal consistency metric (TCM) metric evaluation for multilingual sign corpus.

| Framework | Temporal Consistency Metric (TCM) | | |
|---|---|---|---|
| | RWTH-PHOENIX-Weather 2014T dataset | ISL-CSLTR dataset | How2Sign dataset |
| EBGAN [79] | 0.336 | 0.345 | 0.358 |
| StackGAN [80] | 0.105 | 0.152 | 0.181 |
| BEGAN [81] | 0.191 | 0.162 | 0.185 |
| PROGAN [82] | 0.153 | 0.375 | 0.289 |
| InfoGAN [51] | 0.313 | 0.412 | 0.465 |
| SGAN [83] | 0.424 | 0.411 | 0.421 |
| BiGAN [84] | 0.671 | 0.692 | 0.708 |
| **DynamicGAN** | **0.702** | **0.712** | **0.731** |

the generated results with real data distribution. The results of the FID metric are shown in Table 9.

$$\mathrm{d}^2\left((\mathrm{m_r},\Sigma_\mathrm{r}),(\mathrm{m_f},\Sigma_\mathrm{f})\right) = \|\mathrm{mr}-\mathrm{mf}\|_2^2 + \mathrm{Tr}\left(\Sigma_\mathrm{r}\Sigma_\mathrm{f} - \left(2\left(\Sigma_\mathrm{r}\Sigma_\mathrm{f}\right)^{1/2}\right)\right) \tag{23}$$

We further evaluated our model performance using Temporal Consistency Metric (TCM) metric to provide real score for videos related to consistency in the temporal sequences to produce high quality videos rather than comparing with single frame level. The table 10 list the evaluation scores for TCM Metric and compares with other benchmark datasets.

The user interface based H-DNA implementations are shown in Figures 18 and 19. It shows the sample SL recognition and SL video generation results.

Figure 18 and Figure 19.


**FIGURE 18.** Sample UI based sign gesture recognition results using the H-DNA framework.


**FIGURE 19.** Sample UI application results for video generation using the H-DNA framework.

## V. CONCLUSION

This paper contributes to the development of a deep learning framework for end-to-end sign language recognition, translation, and generation. We addressed the challenges that persist with earlier SL recognition and video generation approaches using the proposed H-DNA framework. We evaluated the model performance using the RWTH-PHOENIX-Weather 2014T dataset, the How2Sign dataset, and the ISL-CSLTR datasets quantitatively and qualitatively. The proposed H-DNA framework is also evaluated qualitatively using various quality metrics. The generated video frames show the quality of the outcome of our work. We achieved a comparatively greater recognition rate and generating performance than earlier approaches. The proposed model has achieved the above 95% classification accuracy towards SL recognition, 38.56 average BLEU score, remarkable human evaluation scores, 3.46 average FID2vid score, 0.921 average SSIM values, 8.4 average Inception Score, 29.73 average PSNR score, 14.06 average FID score, and an average 0.715 TCM Score. These scores are notably higher than earlier models. The evaluation of realism, relevance, and coherence factors is carried out by employing human evaluators and produces good results in real time scenarios.

## FUNDING SUPPORT

This work was financially supported by the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R178), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

## ACKNOWLEDGMENT

The research project was sanctioned by the Science and Engineering Research Board (SERB), India under the Start-up Research Grant (SRG/2019/001338). The authors thank SASTRA Deemed University for providing infrastructural support to conduct the research. They thank all the students for their contribution in collecting the sign videos and the successful completion of the ISL-CSLTR Corpus. And also, they would like to thank Navajeevan, Residential School for the Deaf, College of Spl. D.Ed. and B.Ed., Vocational Centre, and Child Care and Learning Centre, Ayyalurimetta, Nandyal, Andhra Pradesh, India, for their support and contribution.

## REFERENCES

[1] G. Delnevo, R. Girau, C. Ceccarini, and C. Prandi, "A deep learning and social IoT approach for plants disease prediction toward a sustainable agriculture," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7243–7250, May 2021.

[2] I. Siniosoglou, P. Radoglou-Grammatikis, G. Efstathopoulos, P. Fouliras, and P. Sarigiannidis, "A unified deep learning anomaly detection and classification approach for smart grid environments," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1137–1151, Jun. 2021.

[3] G. Vallathan, A. John, C. Thirumalai, S. Mohan, G. Srivastava, and J. C.-W. Lin, "Suspicious activity detection using deep learning in secure assisted living IoT environments," *J. Supercomput.*, vol. 77, no. 4, pp. 3242–3260, Apr. 2021.

[4] S. Ramaswamy and N. DeClerck, "Customer perception analysis using deep learning and NLP," *Proc. Comput. Sci.*, vol. 140, pp. 170–178, Jan. 2018.

[5] V. Pasquadibisceglie, A. Appice, G. Castellano, and D. Malerba, "A multi-view deep learning approach for predictive business process monitoring," *IEEE Trans. Services Comput.*, vol. 15, no. 4, pp. 2382–2395, Jul. 2021.

[6] T. Islam, T. A. Chisty, and A. Chakrabarty, "A deep neural network approach for crop selection and yield prediction in Bangladesh," in *Proc. IEEE Region 10th Hum. Technol. Conf. (R-HTC)*, Dec. 2018, pp. 1–6.

[7] J. Williams, P. Dryburgh, A. Clare, P. Rao, and A. Samal, "Defect detection and monitoring in metal additive manufactured parts through deep learning of spatially resolved acoustic spectroscopy signals," *Smart Sustain. Manuf. Syst.*, vol. 2, no. 1, Nov. 2018, Art. no. 20180035.

[8] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnol.*, vol. 338, pp. 831–838, Aug. 2015.

[9] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, Feb. 2019.

[10] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, "Deep learning detecting fraud in credit card transactions," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, Apr. 2018, pp. 129–134.

[11] P. Bellot, G. de los Campos, and M. Pérez-Enciso, "Can deep learning improve genomic prediction of complex human traits?" *Genetics*, vol. 210, no. 3, pp. 809–819, Nov. 2018.

[12] D. Liciotti, M. Bernardini, L. Romeo, and E. Frontoni, "A sequential deep learning application for recognising human activities in smart Homes," *Neurocomputing*, vol. 396, pp. 501–513, Jul. 2020.

[13] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?" *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015.

[14] Y. Lin, H. Lei, P. Clement Addo, and X. Li, "Machine learned resume-job matching solution," 2016, *arXiv:1607.07657*.

[15] N. J. Cronin, T. Rantalainen, J. P. Ahtiainen, E. Hynynen, and B. Waller, "Markerless 2D kinematic analysis of underwater running: A deep learning approach," *J. Biomech.*, vol. 87, pp. 75–82, Apr. 2019.

[16] W. Zhang, L. Sun, X. Wang, Z. Huang, and B. Li, "SEABIG: A deep learning-based method for location prediction in pedestrian semantic trajectories," *IEEE Access*, vol. 7, pp. 109054–109062, 2019.

[17] R. Elakkiya, "Machine learning based intelligent automated neonatal epileptic seizure detection," *J. Intell. Fuzzy Syst.*, vol. 40, no. 5, pp. 1–9, 2021.

[18] R. Elakkiya, K. S. S. Teja, L. J. Deborah, C. Bisogni, and C. Medaglia, "Imaging based cervical cancer diagnostics using small object detection-generative adversarial networks," *Multimedia Tools Appl.*, vol. 81, pp. 1–17, Jan. 2021.

[19] R. Elakkiya, P. Vijayakumar, and M. Karuppiah, "COVID_SCREENET: COVID-19 screening in chest radiography images using deep transfer stacking," *Inf. Syst. Frontiers*, vol. 23, no. 6, pp. 1–15, 2021.

[20] G. Padmapriya, R. Elakkiya, and M. Prakash, "Deep learning based Parkinson's disease prediction system," in *Machine Learning and IoT for Intelligent Systems and Smart Applications*. Boca Raton, FL, USA: CRC Press, 2021, pp. 97–111.

[21] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Applying deep learning approaches for network traffic prediction," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 2353–2358.

[22] R. N. Babu, V. Sowmya, and K. P. Soman, "Indian car number plate recognition using deep learning," in *Proc. 2nd Int. Conf. Intell. Comput., Instrum. Control Technol. (ICICICT)*, Jul. 2019, pp. 1269–1272.
[23] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.
[24] K. T. P. Nguyen and K. Medjaher, "A new dynamic predictive maintenance framework using deep learning for failure prognostics," *Rel. Eng. Syst. Saf.*, vol. 188, pp. 251–262, Aug. 2019.
[25] K. Orita, K. Sawada, R. Koyama, and Y. Ikegaya, "Deep learning-based quality control of cultured human-induced pluripotent stem cell-derived cardiomyocytes," *J. Pharmacol. Sci.*, vol. 140, no. 4, pp. 313–316, Aug. 2019.
[26] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, and P. Corke, "The limits and potentials of deep learning for robotics," *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 405–420, Apr. 2018.
[27] R. Singh and S. Srivastava, "Stock prediction using deep learning," *Multimedia Tools Appl.*, vol. 7618, pp. 18569–18584, Sep. 2017.
[28] B. Lim and S. Zohren, "Time-series forecasting with deep learning: A survey," *Phil. Trans. Roy. Soc. A*, vol. 379, Apr. 2021, Art. no. 20200209.
[29] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 6, Apr. 2020.
[30] D. Wang, W. Li, X. Liu, N. Li, and C. Zhang, "UAV environmental perception and autonomous obstacle avoidance: A deep learning and depth camera combined solution," *Comput. Electron. Agricult.*, vol. 175, Aug. 2020, Art. no. 105523.
[31] M. Gjoreski, M. Ž Gams, M. Luštrek, P. Genc, J.-U. Garbas, and T. Hassan, "Machine learning and End-to-End deep learning for monitoring driver distractions from physiological and visual signals," *IEEE Access*, vol. 8, pp. 70590–70603, 2020.
[32] P. Hewage, M. Trovati, E. Pereira, and A. Behera, "Deep learning-based effective fine-grained weather forecasting model," *Pattern Anal. Appl.*, vol. 241, pp. 343–366, Feb. 2021.
[33] M. Toğaçar, B. Ergen, and Z. Cömert, "COVID-19 detection using deep learning models to exploit social mimic optimization and structured chest X-ray images using fuzzy color and stacking approaches," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103805.
[34] S. Reddy, N. Srikanth, and G. S. Sharvani, "Development of kid-friendly Youtube access model using deep learning," in *Data Science and Security*. Singapore: Springer, 2021, pp. 243–250.
[35] O. Zavala-Romero, A. L. Breto, I. R. Xu, Y. C. C. Chang, N. Gautney, P. A. Dal, and R. Stoyanova, "Segmentation of prostate and prostate zones using deep learning," *Strahlentherapie und Onkologie*, vol. 19610, pp. 932–942, Oct. 2020.
[36] A. A. Barbhuiya, R. K. Karsh, and R. Jain, "CNN based feature extraction and classification for sign language," *Multimedia Tools Appl.*, vol. 802, pp. 3051–3069, Jan. 2021.
[37] S. Aly and W. Aly, "DeepArSLR: A novel signer-independent deep learning framework for isolated Arabic sign language gestures recognition," *IEEE Access*, vol. 8, pp. 83199–83212, 2020.
[38] C. K. Lee, K. K. Ng, C. H. Chen, H. C. Lau, S. Y. Chung, and T. Tsoi, "American sign language recognition and training method with recurrent neural network," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114403.
[39] Q. Xiao, X. Chang, X. Zhang, and X. Liu, "Multi-information spatial–temporal LSTM fusion continuous sign language neural machine translation," *IEEE Access*, vol. 8, pp. 216718–216728, 2020.
[40] R. Elakkiya, P. Vijayakumar, and N. Kumar, "An optimized generative adversarial network based continuous sign language classification," *Expert Syst. Appl.*, vol. 182, Nov. 2021, Art. no. 115276.
[41] R. Elakkiya, "Machine learning based sign language recognition: A review and its research frontier," *J. Ambient Intell. Hum. Comput.*, vol. 12, no. 7, pp. 1–20, 2020.
[42] R. Elakkiya and B. NATARAJAN, "ISL-CSLTR: Indian sign language dataset for continuous sign language translation and recognition," Mendeley Data Repository, V1, 2021, doi: 10.17632/kcmpdxky7p.1.
[43] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1511–1520.
[44] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," 2016, *arXiv:1606.05328*.
[45] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks," *Int. J. Comput. Vis.*, vol. 1284, pp. 891–908, Apr. 2020.
[46] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 10551–10560.
[47] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[48] Y. Choi, Y. Uh, J. Yoo, and J. W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8188–8197.
[49] S. Tulyakov, M. Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.
[50] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," 2015, *arXiv:1506.05751*.
[51] X. Chen, Y. Duan, R. Houthooft, J. Schulman, and I. A. P. Sutskever, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2016, pp. 2180–2188.
[52] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.
[53] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[54] I. K. Dutta, B. Ghosh, A. Carlson, M. Totaro, and M. Bayoumi, "Generative adversarial networks in security: A survey," in *Proc. 11th IEEE Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2020, pp. 0399–0405.
[55] R. H. Randhawa, N. Aslam, M. Alauthman, H. Rafiq, and F. Comeau, "Security hardening of botnet detectors using generative adversarial networks," *IEEE Access*, vol. 9, pp. 78276–78292, 2021.
[56] Y. Zhu, Y. Zhang, H. Zhang, J. Yang, and Z. Zhao, "Data augmentation of X-ray images in baggage inspection based on generative adversarial networks," *IEEE Access*, vol. 8, pp. 86536–86544, 2020.
[57] R. Sujatha, J. M. Chatterjee, N. Z. Jhanjhi, and S. N. Brohi, "Performance of deep learning vs machine learning in plant leaf disease detection," *Microprocessors Microsyst.*, vol. 80, Feb. 2021, Art. no. 103615.
[58] N. Eldeen M. Khalifa, M. Hamed N. Taha, A. E. Hassanien, and S. Elghamrawy, "Detection of coronavirus (COVID-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest X-ray dataset," 2020, *arXiv:2004.01184*.
[59] B. Espejo-Garcia, N. Mylonas, L. Athanasakos, E. Vali, and S. Fountas, "Combining generative adversarial networks and agricultural transfer learning for weeds identification," *Biosystems Eng.*, vol. 204, pp. 79–89, Apr. 2021.
[60] F. Taymouri, R. M. La, S. Erfani, Z. D. Bozorgi, and I. Verenich, "Predictive business process monitoring via generative adversarial nets: The case of next event prediction," in *Proc. Int. Conf. Bus. Process Manage.* Cham, Switzerland: Springer, 2020, pp. 237–256.
[61] Y. Gu, Y. Peng, and H. Li, "AIDS brain MRIs synthesis via generative adversarial networks based on attention-encoder," in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 629–633.
[62] B. Lütjens, B. Leshchinskiy, C. Requena-Mesa, F. Chishtie, N. Díaz-Rodríguez, O. Boulais, A. Sankaranarayanan, A. Piña, Y. Gal, C. Raïssi, A. Lavin, and D. Newman, "Physically-consistent generative adversarial networks for coastal flood visualization," 2021, *arXiv:2104.04785*.
[63] S. Liu, B. Zhang, Y. Liu, A. Han, H. Shi, T. Guan, and Y. He, "Unpaired stain transfer using pathology-consistent constrained generative adversarial networks," *IEEE Trans. Med. Imag.*, vol. 40, no. 8, pp. 1977–1989, Aug. 2021.
[64] K. Vo, E. K. Naeini, A. Naderi, D. Jilani, A. M. Rahmani, N. Dutt, and H. Cao, "P2E-WGAN: ECG waveform synthesis from PPG with conditional Wasserstein generative adversarial networks," in *Proc. 36th Annu. ACM Symp. Appl. Comput.*, 2021, pp. 1030–1036.
[65] Y. Shi, X. Zhang, Q. Hu, and H. Cheng, "Data recovery algorithm based on generative adversarial networks in crowd sensing Internet of Things," *Personal Ubiquitous Comput.*, vol. 2020, pp. 1–14, Jul. 2020.

[66] C. Davi and U. Braga-Neto, "A semi-supervised generative adversarial network for prediction of genetic disease outcomes," in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2021, pp. 1–6.
[67] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[68] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[69] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[70] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.
[71] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[72] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7784–7793.
[73] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, and X. Giro-i-Nieto, "How2Sign: A large-scale multimodal dataset for continuous American sign language," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2735–2744.
[74] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," 2017, *arXiv:1705.09368*.
[75] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 11303–11312.
[76] T. S. Dias, J. J. A. M. Júnior, and S. F. Pichorim, "An instrumented glove for recognition of Brazilian sign language alphabet," *IEEE Sensors J.*, vol. 22, no. 3, pp. 2518–2529, 2021.
[77] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Syst. Appl.*, vol. 150, Jul. 2020, Art. no. 113336.
[78] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods compariso," in *Proc. IEEE/CVF winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 1459–1469.
[79] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," 2016, *arXiv:1609.03126*.
[80] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5907–5915.
[81] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," 2017, *arXiv:1703.10717*.
[82] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.
[83] A. Odena, "Semi-supervised learning with generative adversarial networks," 2016, *arXiv:1606.01583*.
[84] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," 2016, *arXiv:1605.09782*.
[85] B. Natarajan, R. Elakkiya, and M. L. Prasad, "Sentence2SignGesture: A hybrid neural machine translation network for sign language video generation," *J. Ambient Intell. Hum. Comput.*, vol. 2022, pp. 1–15, Jan. 2022.
[86] B. Natarajan and R. Elakkiya, "Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks," *Soft Comput.*, vol. 2022, pp. 1–23, Jun. 2022.

**B. NATARAJAN** received the Bachelor of Engineering degree, in 2011, and the Master of Engineering degree, in 2015. He is currently pursuing the Ph.D. degree with the School of Computing, SASTRA University, Thanjavur. He has seven years of teaching experience and has published many articles in leading international journals. His research interests include computer vision, machine learning, deep learning, and sign language development.

**E. RAJALAKSHMI** received the B.E. degree in information technology from the Cummins College of Engineering, Pune, in 2018, and the M.Tech. degree in computer science and engineering from SASTRA Deemed University, Thanjavur, in 2020, where she is currently pursuing the Ph.D. degree.

She is currently working as a Project Associate with SASTRA Deemed University. Her research interests include sign language recognition, music emotion recognition, deep neural network, image processing, and computer vision. She has contributed various articles and chapters for many high-quality Scopus and SCI/SCIE indexed journals, conferences, and books. She is a Lifetime Member of International Association of Engineers and a member of Association for Computing Machinery.

**R. ELAKKIYA** received the Doctor of Philosophy from Anna University, Chennai, in 2018. She is currently working as an Assistant Professor with the Department of Computer Science and Engineering, School of Computing, SASTRA University, Thanjavur. She got three patents. She has published more than 20 research papers in leading journals, conference proceedings, and book including IEEE, Elsevier, and Springer. She is currently an Editor of Information Engineering and Applied Computing journal and also a Life Time Member of International Association of Engineers.

**KETAN KOTECHA** is currently an Administrator and a Teacher of deep learning with Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune. He has expertise and experience in cutting-edge research and projects in A.I. and deep learning for the last 25 years. He has published more than 100 widely in several excellent peer-reviewed journals on various topics ranging from cutting edge A.I., education policies, teaching-learning practices, and A.I. for all. He has published three patents and delivered keynote speeches at various national and international forums, including at the Machine Intelligence Laboratory, USA, IIT Bombay under the World Bank Project, the International Indian Science Festival organized by the Department of Science and Technology, Government of India, and many more. His research interests include artificial intelligence, computer algorithms, machine learning, and deep learning. He was a recipient of the two SPARC Projects worth INR 166 lakhs from MHRD Government of India in A.I. in collaboration with Arizona State University, USA, and The University of Queensland, Australia. He was also a Recipient of numerous prestigious awards, such as Erasmus+ Faculty Mobility Grant to Poland, the DUO-India Professors Fellowship for research in responsible A.I. in collaboration with Brunel University, U.K., the LEAP Grant at Cambridge University, U.K., the UKIERI Grant with Aston University, U.K., and a Grant from the Royal Academy of Engineering, U.K., under Newton Bhabha Fund. He is an Associate Editor of IEEE Access journal.

**AJITH ABRAHAM** (Senior Member, IEEE) received the Master of Science degree from Nanyang Technological University, Singapore, in 1998, and the Ph.D. degree in computer science from Monash University, Melbourne, Australia, in 2001. He is currently the Director of the Machine Intelligence Research Laboratories (MIR Laboratories), a Not-for-Profit Scientific Network for Innovation and Research Excellence Connecting Industry and Academia. The Network with HQ in Seattle, USA, is currently more than 1,500 scientific members from over 105 countries. As an Investigator/a Co-Investigator, he has won research grants worth over more than U.S. $100 Million. Currently, he holds two university professorial appointments. He works as a Professor in artificial intelligence at Innopolis University, Russia, and the Yayasan Tun Ismail Mohamed Ali Professorial Chair in Artificial Intelligence at UCSI, Malaysia. He works in a multi-disciplinary environment. He has authored/coauthored more than 1,400 research publications out of which there are more than 100 books covering various aspects of computer science. One of his books was translated into Japanese and a few other articles were translated into Russian and Chinese. He has more than 46,000 academic citations (H-index of more than 102 as Per Google Scholar). He has given more than 150 plenary lectures and conference tutorials (in more than 20 countries). He was the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over more than 200 members), from 2008 to 2021, and served as a Distinguished Lecturer of IEEE Computer Society representing Europe (2011–2013). He was the Editor-in-Chief of Engineering Applications of Artificial Intelligence (EAAI), from 2016 to 2021, and serves/served on the editorial board for over 15 international journals indexed by Thomson ISI.

**LUBNA ABDELKAREIM GABRALLA** received the B.S.C. and M.Sc. degrees in computer science from the University of Khartoum, and the Ph.D. degree in computer science from the Sudan University of Science and Technology, Khartoum, Sudan. She is currently an Associate Professor with the Department of Computer Science and Information Technology, Princess Nourah Bint Abdulrahman University, Saudi Arabia. Her current research interests include soft computing, machine learning, and deep learning. She became a Senior Fellow (SFHEA), in 2021.

**V. SUBRAMANIYASWAMY** received the B.E. degree in computer science and engineering and the M.Tech. degree in information technology from Bharathidasan University, India, and Sathyabama University, India, and the Ph.D. degree from Anna University, India, and continued the extension work with the Department of Science and Technology support as a Young Scientist Award Holder. He is currently working as a Professor with the SASTRA Deemed University, Thanjavur, India. In total, he has 18 years of experience in academia. He has contributed more than 160 papers and chapters for many high-quality Scopus and SCI/SCIE indexed journals and books. He is on the reviewer board of several international journals and has been a program committee member for several international/national conferences and workshops. He also serves as a guest editor for various special issues of reputed international journals. He is serving as a research supervisor and also a visiting expert to various universities in India. His technical competencies lie in recommender systems, social networks, the Internet of Things, information security, and big data analytics.