IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 June 2024, accepted 20 June 2024, date of publication 28 June 2024, date of current version 9 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3420776

RESEARCH ARTICLE

Hybrid InceptionNet Based Enhanced Architecture for Isolated Sign Language Recognition

DEEP R. KOTHADIYA[1], CHINTAN M. BHATT[2], (Senior Member, IEEE), HENA KHARWA[1], AND FELIX ALBU[3], (Senior Member, IEEE)

[1]U & P U Patel Department of Computer Engineering, Faculty of Technology (FTE), Chandubhai S. Patel Institute of Technology (CSPIT), Charotar University of Science and Technology (CHARUSAT), Changa 388421, India
[2]Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar 382007, India
[3]Department of Electronics, Valahia University of Târgoviște, 130004 Târgoviște, Romania

Corresponding authors: Chintan M. Bhatt (chintan.bhatt@sot.pdpu.ac.in) and Felix Albu (felix.albu@valahia.ro)

The work of Felix Albu was supported in part by the Romanian Ministry of Research, Innovation and Digitization; in part by the National Scientific Research Council (CNCS); in part by the Executive Unit for Financing Higher Education, Research, Development and Innovation (UEFISCDI) under Project PN-III-P4-PCE-2021-0780; and in part by the National Research-Development and Innovation Plan (PNCDI).

ABSTRACT Sign language is a common way of communication for people with hearing and/or speaking impairments. AI-based automatic systems for sign language recognition are very desirable since they can reduce barriers between people and improve Human-Computer Interaction (HCI) for the impaired community. Automatically recognizing sign language is still an open challenge since the sign language itself has a complex structure to convey messages. The key role is played by the isolated signs that refer to single gestures carried out by hand movements. In the last decade, research has improved the automatic recognition of isolated sign language from videos using machine learning approaches. Starting from a comprehensive analysis of existing recognition techniques, with an in-depth focus on existing public datasets, the study proposes an advanced convolution-based hybrid Inception architecture to improve the recognition accuracy of isolated signs. The main contributions are to enhance InceptionV4 with optimized backpropagation through uniform connections. Besides, an ensemble learning framework with different Convolution Neural Networks has been also introduced and exploited to further increase the recognition accuracy and robustness of isolated sign language recognition systems. The effectiveness of the proposed learning approaches has been proved on a benchmark dataset of isolated sign language gestures. The experimental results demonstrate that the proposed ensemble model outperforms sign identification, yielding higher recognition accuracy (98.46%) and improved robustness.

INDEX TERMS Sign language recognition, gesture recognition, isolated sign, deep learning, computer vision.

I. INTRODUCTION

In this world where communication knows no bounds, sign language bridges the gap between the hearing impaired, enabling the exchange of ideas and emotions in a visually captivating manner [1]. However, in this era of advanced technology, the recognition of signs can reduce the communication gap for impaired communities and the rest of the world [2]. Sign language is a gesture-based medium of communication and can be divided into two basic categories: static and dynamic. Static sign language mainly consists of digits, an alphabet, and some common words, while dynamic sign language is a complete foam of sign language to communicate proper meaning. The input modality of static signs is an image, while dynamic signs have videos. Dynamic signs are also divided into two categories: isolated and continuous signs. Isolated sign gestures are used to express words, whereas continuous sign gesture is a

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

combination of isolated signs to express syntax meaningfully. So, isolated sign gestures are the backbone of sign language [3].

The sign language recognition system is an essential AI-driven system to connect the impaired community to the rest of the world. Sign language recognition systems not only reduce communication barriers but also improve Human-Computer Interaction (HCI) systems for the impaired community [4]. The recognition of isolated sign language improves communication with the rest of the community and the usability of automation systems. The isolated sign language recognition system represents a remarkable fusion of cutting-edge artificial intelligence, computer vision, and machine learning technologies [5]. It is designed with a singular purpose to comprehend and interpret the intricate movements and gestures of sign language users, transforming their expressive motions into meaningful and actionable information. Unlike traditional static recognition systems that focus on individual signs, this revolutionary technology possesses the ability to perceive the fluidity and context within sign language conversations. It deciphers the dynamic interplay of gestures, facial expressions, and body language, enabling a deeper comprehension of the linguistic nuances and emotional nuances embedded within each interaction [6]. Computer vision is a branch of Deep Learning (DL) that allows systems and smart devices to retrieve useful information from multimedia contexts like images, videos, and other visual inputs, and then recommend actions based on that information. Deep learning models enable computers to think, computer vision enables them to view, evaluate, and interpret their surroundings [7]. DL algorithms have accelerated progress in a wide range of fields, including speech recognition, visual object recognition, video categorization, and even medicine development. A deep learning model includes several processing layers that allow it to learn high-level representations on its own. DL does not need considerable feature engineering or concept knowledge [8]. Furthermore, with so many deep transformations, it can learn extremely complex functions and deal with difficult classification and identification problems. As a result, deep learning has advanced in numerous areas, including isolated sign recognition. We examine many deep learning models and examine their distinct advantages for such assignments. Table 1 demonstrates an analysis of different deep-learning approaches that can be used for isolated sign-language recognition [9].

The main contributions of the proposed research work are: i) A comprehensive study and analysis of different approaches have been carried out for isolated sign language recognition; ii) Using hybrid InceptionNet architecture, this study proposes an efficient deep learning architecture for isolated sign language recognition; iii) Simulation of the proposed study demonstrates comparative analysis with SOTA deep learning models and explores the effectiveness of different combinational studies for digitalization of isolated sign language.

TABLE 1. Advantage of deep learning models for sign language recognition.

Models	Description and advantages
CNN	Multiple convolutional layers capture the spatial connection and are frequently used as excellently targeted feature extractors.
RNN	When investigating the temporal connection in data, variations such as LSTM are frequently used.
Auto-encoder	A neural network that is fed forward that learns uncontrolled deep features
Hybrid deep models	The combination of some deep models, based on each model's power, to achieve superior performance
Transfer Learning	Pre-train models are again trained on a targeted dataset, which helps to enhance feature learning.

The rest of the paper is organized as follows: Section II demonstrates the previous work done in the field of isolated sign language recognition and analyzes the research gap that can be solved by the proposed methodology. Section III represents a detailed analysis of the proposed methodology, and Section IV demonstrates the dataset and result analysis of the proposed architecture with various benchmark isolated datasets.

II. RELATED WORK

The different approaches and methods available for sign language recognition are summarized in this section. Effective methods to recognize isolated signs are glove-based sign gesture recognition [12]. They make use of gloves equipped with sensors that detect finger movements and hand gestures can be used to capture sign language gestures directly from the user's hand, in a wearable method. Data from multiple sensors on the glove can be combined to get a comprehensive representation of the gesture in the sensor fusion approach. The biggest drawback of these approaches is the dependency on gloves. The system may not work without a specific type of gloves [13].

Recognition of sign language based on computer vision utilized techniques such as depth sensors, color cameras, pose estimation, and feature extraction [10]. Depth cameras like Microsoft Kinect or Intel RealSense can capture the depth information of the scene, which helps in capturing the 3D structure of sign gestures. Computer vision techniques like pose estimation can be used to extract joint positions and movements from the captured images or video frames. Various features such as edges, corners, and texture can be extracted from the images to represent the sign gestures. The computer vision-based approach uses different machine learning and deep learning approaches to improve the performance of sign language recognition systems [11].

Pose estimation plays a significant role in sign language recognition by capturing the positions and orientations of key points on the signer's body, particularly the hands and sometimes the face. Use a pose estimation model, such as OpenPose or PoseNet, to extract the 2D or 3D

coordinates of relevant body parts (joints) from each frame of the video [14]. Focus on capturing hand and arm movements. For 3D pose estimation, depth cameras or multiple calibrated cameras can be used to capture the 3D positions of body joints. Pose estimation-based methods used the identification of non-manual gesture components to improve recognition accuracy for sign language. There are basically two approaches used for gesture-based sign language recognition: one should be text-based, which has a high error rate and requires a large vocabulary. The second method is vision-based, which uses images and video to represent the words. In the following, the most relevant works using computer vision for sign language recognition are listed and discussed [15].

Fink et al. [16] created a recognition system based on sign language. They made use of a dataset collecting frames that display only one sign. Each video was then preprocessed by at first reducing the frame width to 270 pixels while maintaining its aspect ratio. The footage was then cut into 50 frames and finally, patches of 224 × 224 pixels were cut at random from the video and sent to networks. A hybrid architecture consisting of VGG, LSTM and C3D was exploited for sign language recognition. However, the proposed technique lies on a complex architecture made of convolution and recurrent layers to achieve a top-one accuracy not greater than 51.5%.

De Coster et al. [17] pioneered the development of automated sign language recognition. It is made up of 36,302 samples from 226 different sign categories. From RGB video data, Pre-processed multi-modal input is extracted. In this approach, posture flow, an optical flow-inspired approach for depicting body movements based on key points in posture, is introduced. All samples are sent as individual RGB and depth video files with 512 by 512 pixels, temporal resolution of 30 frames per second (FPS), and spatial resolution of 30 FPS. For every frame, the ResNet-34 network is utilized to extract a 512-dimensional feature vector. VTN-PF, VTN-HC, and VTN categorization systems are utilized. This approach has 92.92% accuracy.

Hao et al. [18] proposed a continuous sign language recognition and self-mutual distillation learning system. The frames in both datasets are scaled to 256 256 and then cropped to 224 224. During training, we enrich the data using a random crop and a 50% horizontal flip. The visual module converts short-term spatial-temporal information into visual characteristics for each input sequence. For feature extraction techniques, 2D CNN and 1D CNN were utilized, which encode spatial and short-term temporal information, respectively. CNN+LSTM+HMM, FCN, and STMC are the classification algorithms utilized to build the SMDK system. Even though a hybrid convolution and recurrent approach is used, it leads to a higher WER of 20.8%.

Boháček and Hrúz [19] developed a sign language recognition system at word-level based on transformers. The Vision API is used for preprocessing video frames to assess posture (head, body, and hand landmarks). There were 54 body landmarks, 5 head landmarks, and 21 hand landmarks discovered. To train the model over 350 learning cycles, an SGD optimizer with an initial learning rate of 0.0001 is utilized. Authors have extracted the picture from the tape, identified all-important body landmarks, augmented it, and normalized it. In this manner, the authors analyzed each frame of the recording to determine what should be sent into the transformer model. TK-3D ConvNet, Fusion-3, and GCN-BERT are the classification methods employed. Although transformer-based encoder architecture leads to increased model parameters and can archive moderate top 1 accuracy as 63.18% (WLAZL) and 43.78% (LSA64).

Campos-Taberner et al. [20] proposed an American Sign Language word recognition technique based on skeletal video. The weighted least squares (WLS) algorithm is used for preprocessing to minimize noise in 3D skeletal video sequences; the EKF method is used to follow deep hand motion trajectories across several video frames; MIC is used for sturdy feature selection; features are scaled to regulate hand motions and accommodate new signers; as well as skeletal video frames are corrected to regulate the beginning frame coordinates and the position of all subsequent frames. At each video frame, to scale independent characteristics, the Z-score transformation is applied, within a certain threshold range. The classification algorithm utilized was Multistack Bi-LSTM, which has a 97.98% accuracy.

Armagan et al. [21] used ensemble learning to create a system for isolated sign language recognition. The authors have collected several modalities of data from the original RGB photos during the preprocessing step. For estimating (x,y) as 2D information about the positions of joints and significant places on the face in a specific RGB image, the OpenPose and MMpose frameworks were utilized. The BODY model is utilized for OpenPose, this results in 18 body joints, 21 points for each single hand, and 69 points for the face. MMPose contains 68 facial points, 21 hand points, and 23 body joints (DARK with HRNet backbone. Ensemble techniques employing I3D, TimeSformer, and SPOTER classification models were utilized for the ensemble technique, and the accuracy was 73.84%. the complex pose estimation-based transform architecture is still not able to archive acceptable accuracy for isolated sign language recognition.

Chen et al. [22] created a 3D Convolutional Network with Multi-Scale Attention convolutional network to identify multimodal gestures. Using the equidistant sampling with the random jitter technique, videos were preprocessed with 16 frames per sign gesture for the proposed IsoGD and 32 frames per sign gesture for Briareo dataset. During training, the video sample's frames were all arbitrarily chopped to 224. Frames had been center-cropped to the identical 224 × 224 size during the inference phase. The starting learning rate was set at 0.01, and it was doubled every three epochs by 0.1. The categorization techniques for gesture recognition are as follows: 3D

TABLE 2. A comparative analysis of recent work on isolated sign language using vision-based deep learning models, (WER = Word Error Rate).

Author	Year	Methods	Dataset	Results
J. Fink et al. [16],	2021	VGG+LSTM, C3D , I3D	LSFB and MS-ASL	51.5%, Top 1 accuracy
M. De Coster et al [17],	2021	VTN, VTN-HC, VTN-PF	AUTSL	92.92%
A. Hao et al. [18]	2021	Self-Mutual Knowledge Distillation	PHOENIX14 , PHOENIX14-T	20.8 (WER)
Boháček, M. and Hrúz, M [19]	2022	TK-3D ConvNet , Fusion-3 , GCN-BERT	WLASL and LSA64 datasets	63.18% (WLASL), 43.78% (LSA64),Top 1 accuracy
Campos-Taberner et al. [20]	2020	Multistack Bi-LSTM	SHREC, LMDHG	97.98%
Armagan et al. [21],	2022	Ensemble using I3D, TimeSformer , SPOTER	AUTSL and WLAN SL300	73.84%
H. Chen et al. [22],	2022	3D Convolutional Network with Multi-Scale Attention	Chalearn LAP Isolated gesture and Briareo	68.15%
H. Chen et al. [23]	2024	SignVTCL (Multi Model)	Phoenix-2014	17.3(WER)%
K. M. Hama Rawf et al. [24]	2024	CNN	Kurdish Sign Language (KuSL)	99.05%
Al Khuzayem, L et al. [25]	2024	2D-CNN	Arabian Sign Language	94.79%
V. Singla, et al. [26]	2024	Visual Transformers	Indian Sign Language	97.52%
J. M. Joshi & D. U. Patel [27]	2024	LSTM	Indian Sign Language (Gujarati)	98.45%

Convolutional Network with Multi-Scale Attention, with 68.15%
accuracy.

Table 2 shows a comparative analysis of various recent sign language recognition models and methods. Analysis finds that many recognition systems use ensemble approaches with convolution followed by recurrent networks [16, 18, 20], which leads to more parameters and consequently increases computational time. Many methodologies have used special cameras like the MS Kinect to collect additional information for depth and RBG values [14,19, 22]. To alter the dependencies on special cameras, authors have proposed a hybrid convolution-based architecture for isolated sign language recognition with minimal data training cycles.

III. MATERIALS AND METHODS

The proposed study uses an ensemble learning methodology with a vanilla CNN model and a modified Inceptionv4 network to recognize sign gestures from video. The proposed methodology uses the stem and inception layers, which are used to make random projections on spatial and temporal features extracted from the convolution blocks, with an additional weighting layer controlling the strength of the

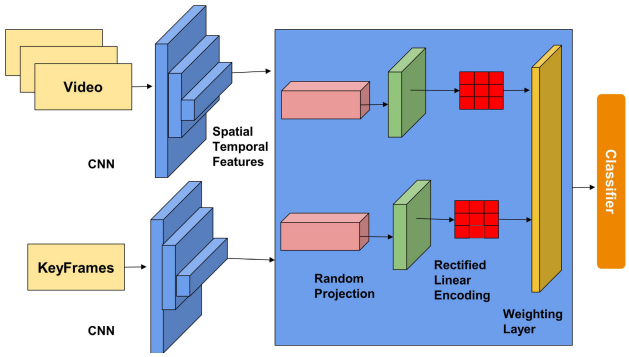

FIGURE 1. The architecture of the proposed InceptionNet-based model for sign language.

input fed to the model. The stem layer was utilized to extract complex features from the input being fed to the fully connected layer to generate predictions with the classifier to help our model create complicated predictions. The rest of the layers combine with sophisticated features of recognizing higher-level patterns from input isolated sign gestures. Inception layers were used to apply various filter sizes concurrently and concatenate the resultant feature maps [28]. This technique enables networks to learn both local and global properties at various sizes, as well as capture a more diversified range of patterns from isolated signs. The pooling and normalization layers minimize the dimensionality of the feature maps, which helps to increase the network's efficiency in terms of computation and memory usage. The proposed methodology improves the prediction performance and efficiency of isolated sign language recognition. Figure 1 illustrates the proposed hybrid architecture for isolated sign language.

A. ENSEMBLE LEARNING

The ensemble learning approach is used to improve prediction accuracy by associating different deep learning models. Authors have proposed an ensemble learning architecture for isolated sign language recognition in which vanilla CNN within InceptionNet models are used to modify the trade-off between reasoning complexity and accuracy at runtime. The proposed method uses two networks of varying sizes for learning layers at inference time. The weighted average ensemble method has been used to cumulate features from the Inception and Convolution networks.

The proposed study employs the Inception v4 deep learning architecture for isolated Indian sign language recognition. Compared to a typical convolution network, the proposed network is deeper, having six convolution layers followed by six maximum pooling layers, followed by flattening, and a dense layer. A dropout of 0.3 was used for the fully connected layers to avoid overfitting the training data. The first convolutional layer is followed by the downsampling of the input frames in the initial input layer, which is referred to as a pre-processing layer for input video. 64 kernels of size

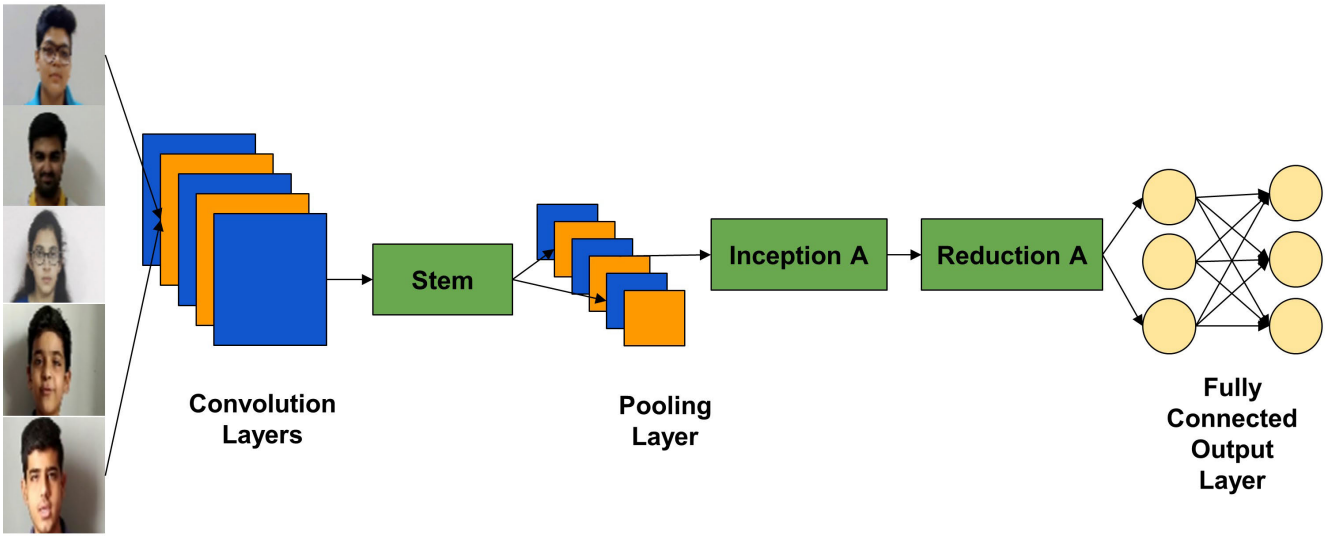

FIGURE 2. Conceptual architecture of proposed convolutional networks.

(3×3) were used for the remaining convolutional layers, and 32 kernels of size (3×3) were used for the rest of the layers. Convolutional feature maps from the first, second, and fifth layers are merged with 33. The classification probabilities used in the last output classification fully connected layer, this feature vector was assigned to soft-max activation and fully connected layers (FC). Figure 2 illustrates the functional architecture of the proposed InceptionNet module.

Convolutional layers are based on filters, which function like the weights of the Completely Connected Network. The kernel moves over frames, resulting in the output known as a feature map. We conducted matrix multiplication at each position on the input before integrating the result. Equation 1 [29] defines mathematical formulation for output feature map.

$$N_x^r = \frac{N_x^{r-1} - L_x^r}{S_x^r} + 1; N_y^r = \frac{N_y^{r-1} - L_y^r}{S_y^r} + 1 \quad (1)$$

where N_x, N_y represents the width and height of the previous layer's output feature map, L_x, L_yrepresents the kernel size, and $S_x, S_y$the number of pixels skipped by the kernel in horizontal and vertical directions, while r is the layers. Convolution was applied to a kernel, and the input feature map was utilized to generate the output feature map, as formulated in equation 2 [30].

$$x_1(m, n) = (j * r)(m, n) \quad (2)$$

where $x^1(m.n)$ is a two-dimensional output feature map produced by convolving the L_x, L_y dimensional kernel r. Equation 3 [30] expresses the convolution operation.

$$x^1(m, n) = \sum_{q=-L_y/2}^{q=+L_y/2} j(m - p, n - q) * r(p, q) \quad (3)$$

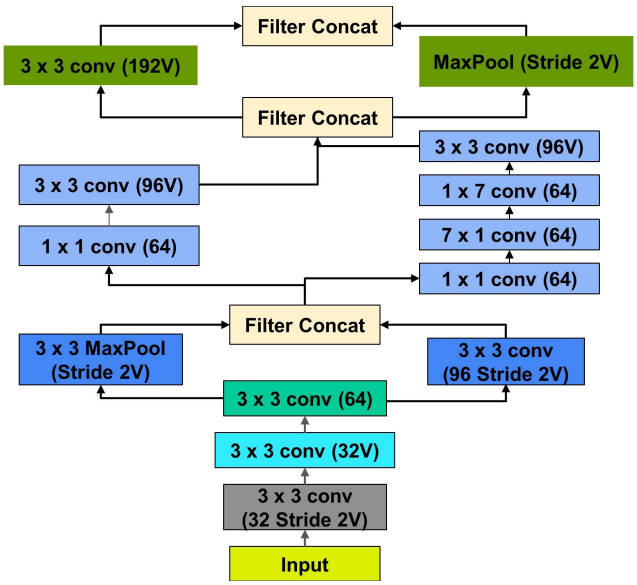

FIGURE 3. Composition of the stem layer.

B. INCEPTION V4 AND STEM LAYERS

In the proposed architecture, the stem module is the first component of the network and is responsible for processing the isolated sign frames and extracting useful features. It is designed with convolutional layers, pooling, and normalization layers used to decrease the geographic input dimensions of the frames. The stem module's output is then transmitted to Inception v4 network, which further processes the extracted weighted matrixes and performs the feature enhancing [31]. The proposed stem composition used in the methodology is illustrated in figure 3.

The major modifications incorporated with Inception v4 to optimize feature learning are i) the use of smaller convolution

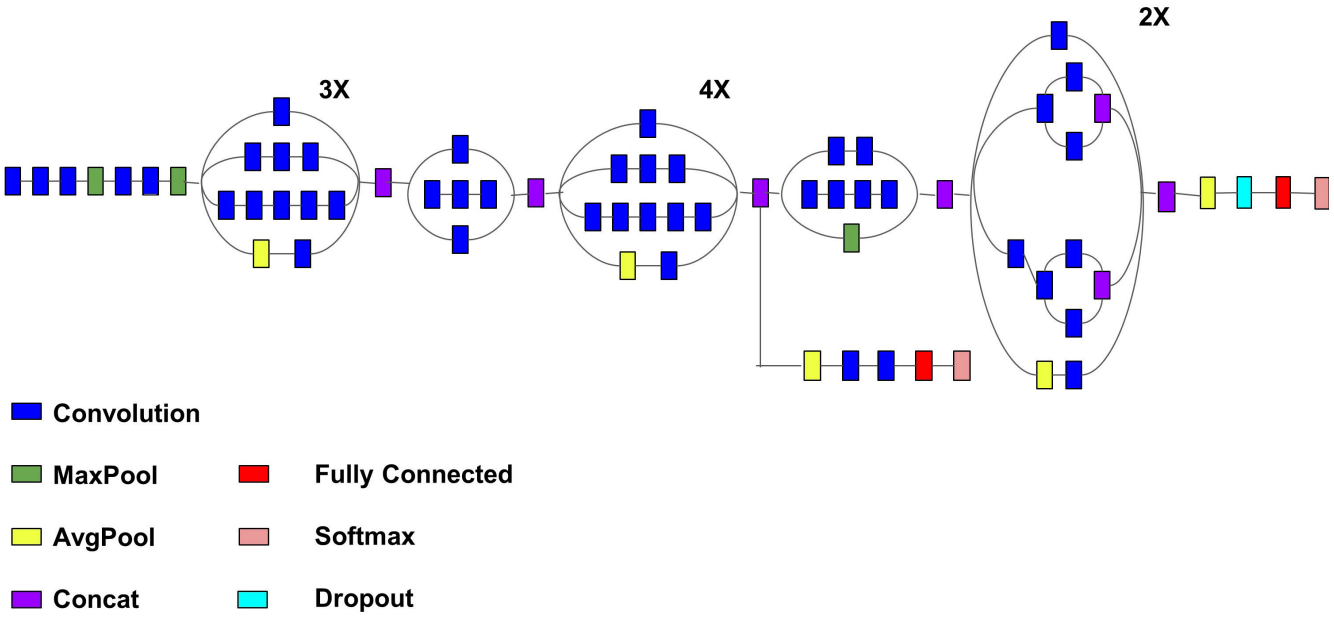

FIGURE 4. The proposed architecture of optimized Inception v4 for isolated sign language.

factorization; ii) Spatial factorization into asymmetric convolutions; iii) the use of auxiliary classifiers; and iv) Efficient grid size reduction. Figure 4 illustrates the proposed network, which consists of 42 layers and an optimized version of the Inception v4.

1) INCEPTION LAYER

The Inception v4 module is an essential part of the proposed sign recognition model. It is intended to capture characteristics from the input frames at various scales and resolutions. Convolutional layer extract features information from the input by using various filter sizes (1×1, 3×3, and 5×5). Each filter size's output is combined with the channel dimension to yield a singular output tensor. This method enables the network to record characteristics at various spatial scales and resolutions of signs, which can be useful for recognizing signs of various sizes and forms in an image. Inceptionv4 also enhances network speed and efficiency with features such as batch normalization and factorized convolutions [32]. Overall, the inception layer allows the network to collect a broad variety of characteristics from input signs, which can contribute to improved sign categorization performance [33]. The full configuration of Inceptionv4 is summarized in Figure 5.

2) RECTIFIED LINEAR UNIT LAYER

ReLU activation function was used in the proposed convolution layer to increase their strength by rendering them non-linear. The main purpose of the activation function is to provide the neural network with nonlinear expression capability, allowing it to better match the findings and enhance accuracy. Due to its linear behavior and computational

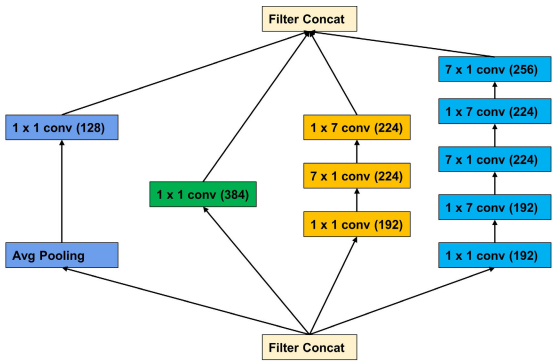

FIGURE 5. Configuration of Inception V4 layers used in proposed model.

simplicity, we ended up feeding RELU as an activation function to our model. The functionality of RELU activation can be calculated as equation 4 [34].

$$f(x) = max(0, x) \tag{4}$$

C. ACTIVATION AND POOLING LAYER

Following each convolution layer in the proposed architecture was the pooling layer. The purpose of deploying the max pooling layer is to lower the spatial size of the convolved features and help reduce overfitting by providing an abstract representation of them. It is defined as a process where the kernel extracts the maximum value of the area it convolves. The representation of features extracted by the convolution layer and polling function is illustrated in Fig. 6.

The softmax activation function is incorporated in our proposed architecture after the fully connected dense layer. Softmax is applied on the top of the retrieved features. The

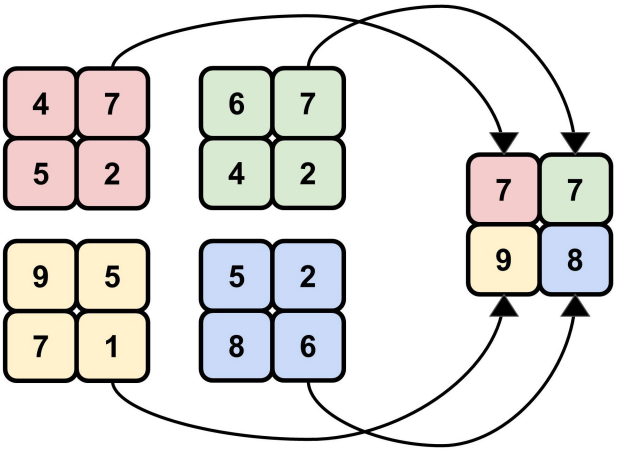

FIGURE 6. Representation of features fetched by the convolution layer and maximum pooling layer.

output of the trainable layer is supplied to the soft-max layer for multiclass classification, which aids in determining the classification probabilities of the input sign. The final classification layer then uses these probabilities to categorize the frames into distinct classes. The softmax function can be evaluated as equation 5 [35]. All values are the elements of the softmax function's input vector, and they can be any real value, positive, zero, or negative, while K is the number of classes.

$$\sigma(\hat{Z}) = \frac{e^{z_i}}{\sum_{i=1}^{k} e^{z_i}} \tag{5}$$

IV. RESULTS

A. DATASET

The dataset utilized for this research effort comprises several sorts of films that may be classified as follows: (i) Bye (ii) Good (iii) Hello (iv) House (v) Morning (vi) Nice (vii) No (viii) Thank You (ix) Welcome (x) Work (xi) Yes. The dataset includes 15 videos belonging to each class. The frame resolution of each video is 1920 × 1080, 30 frames per second, and MPEG-4 encoding. We have divided our dataset into train and test with 0.2 as a split ratio. Figure 7 displays a glimpse of the Isolated Indian sign language dataset (IISL-2020) [36].

B. SIMULATION

The proposed network has a total of 42 layers with 64 channels, which is more than pre-trained CNN, and is made up of 6 convolutional layers followed by pooling and activation layers with a filter size of 3 × 3. A fine-tuned CNN model is combined with Inception v4 layers, followed by a dense dropout and a fully connected layer in the ensemble technique. Ensemble learning models have the benefit of being able to generate better forecasts and improved performance superior to any single contributing model, and their resilience minimizes the spread of predictions and

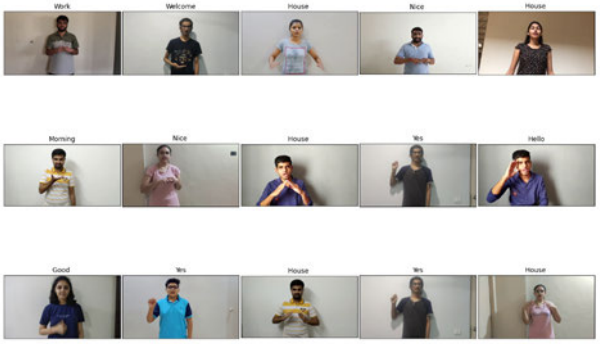

FIGURE 7. Samples from our dataset from 11 classes.

TABLE 3. Units for magnetic properties.

Hyperparameters	Value
Dropout	0.3
Epochs	15
Activation	ReLU
Regularization	Batch Normalization
Optimizer	Adam
Learning Rate	0.0001

model performance. Initially, the specifications and values of Hyperparameters are listed in Table 3.

This residual Convolution neural network includes 42 layers and 64 channels. The first is a convolutional layer with a kernel size of 7 and a stride of 2. The stemming and inception v4 phases are followed by the convolution block with filter sizes of 3 × 3, 3 × 3, and 1 × 1 for 64 channels. Each block is followed by the convolution layer, which comprises 64, 64, and 256 channels. The proposed methodology was evaluated using precision, recall, and F1-rating [37], which are mathematically formulated as equations 6 to 8.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{8}$$

C. RESULTS AND ANALYSIS

Figure 8 depicts the proposed methodology's accuracy and loss graph for the Indian sign language dataset, which has achieved 98.46% accuracy. Proposed method has also accomplished remarkable performance over other isolated sign language datasets like AUTSL [38], DEVISIGN [39], and GSL [40]. Table 4 exhibits the proposed methodology's performance analysis over different datasets.

The proposed ensemble learning model having InceptionV4 finds better performance with stream and reduction layers. Additionally, authors have simulated the proposed architecture with different dropout (dr) and learning rates (lr). Comparative analysis of different isolated sign datasets has been demonstrated in table 5. Authors have also

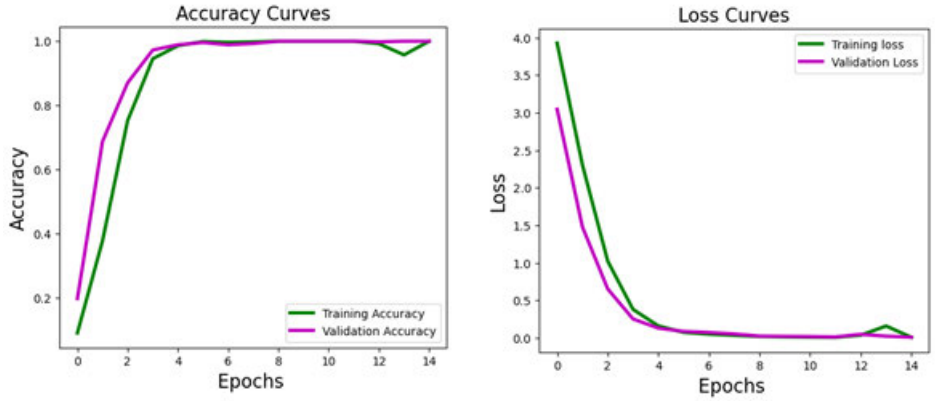

FIGURE 8. Accuracy and loss graphs of the proposed methodology.

TABLE 4. Comparative analysis over different isolated sign datasets.

Dataset	Precision	Recall	F1-rating	Accuracy
IISL	0.99	0.98	0.98	98.46%
AUTSL	0.91	0.95	0.93	94.29%
GSL	0.97	0.96	0.96	97.01%
DEVSIGN	0.95	0.96	0.95	96.55%

TABLE 5. Comparative accuracy analysis of the proposed methodology over different dropout and learning rates.

Dataset	lr=0.001		lr = 0.0001	
	Dropout=0.3	Dropout=0.2	Dropout=0.3	Dropout=0.2
IISL	84.22	89.29	98.46	96.31
AUTSL	80.03	85.16	94.29	91.69
GSL	90.21	91.05	97.01	95.44

TABLE 6. Comparative analysis of proposed model with SOTA classifier of deep learning.

Model	Dataset	Accuracy (20 epoch)	Accuracy (25 epoch)
ANN	IISL	84.01%	85.54%
	AUTSL	79.84%	81.35%
	GSL	77.73%	77.68%
RF	IISL	74.39%	76.02%
	AUTSL	69.10%	73.87%
	GSL	69.89%	71.33%
KNN	IISL	64.58%	65.09%
	AUTSL	65.37%	68.91%
	GSL	65.94%	65.29%
SVM	IISL	74.26%	78.31%
	AUTSL	76.11%	77.48%
	GSL	75.39%	76.09%
MLP	IISL	86.40%	89.61%
	AUTSL	84.70%	85.55%
	GSL	88.61%	89.14%

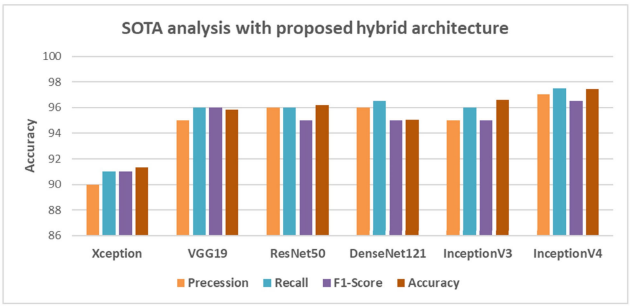

FIGURE 9. Comparative analysis of IISL-2020 over different deep learning models with proposed hybrid architecture.

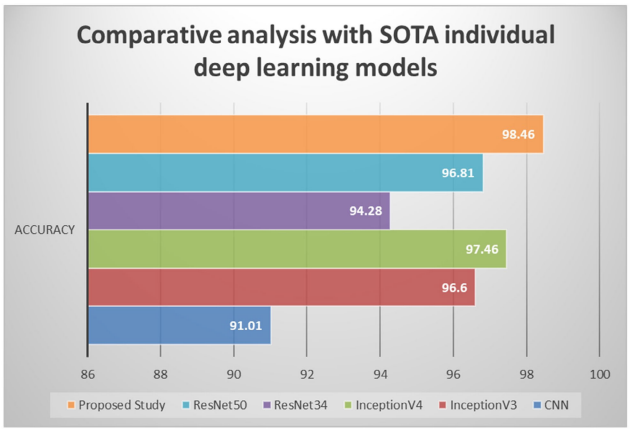

FIGURE 10. Comparative analysis of proposed ensemble architecture with individual deep learning models.

experimented proposed architecture with different classifier and simulated with different benchmark dataset to validate efficacy of proposed architecture demonstrated in table 6.

Authors have also simulated and analyzed the performance of other deep learning models like Xception [41], VGG16 [42], ResNet50 [43], DenseNet121 [44], and Inception [45], [46]. Figure 9 demonstrates the ability of the proposed hybrid (Vanilla CNN + variants) to find the best performance for isolated sign language recognition. The authors also compare the vanilla inception model with InceptionV4 and the proposed ensemble InceptionV4. Figure 10 demonstrates the effectiveness of proposed ensemble approach by comparing individual SOTA deep learning models for recognition of sign language. Figure 11 illustrates a comparative analysis of different variants of inception net with the proposed hybrid

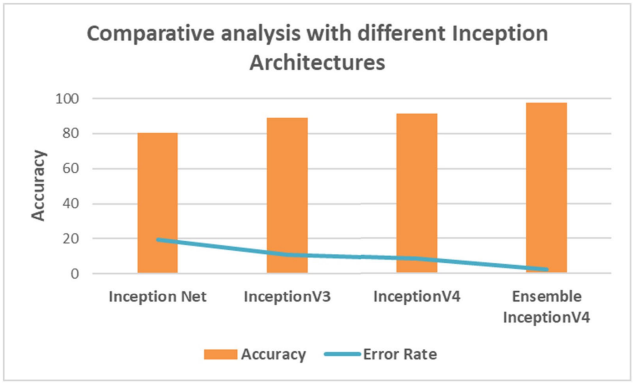

FIGURE 11. Comparative analysis with different variants of Inception networks.

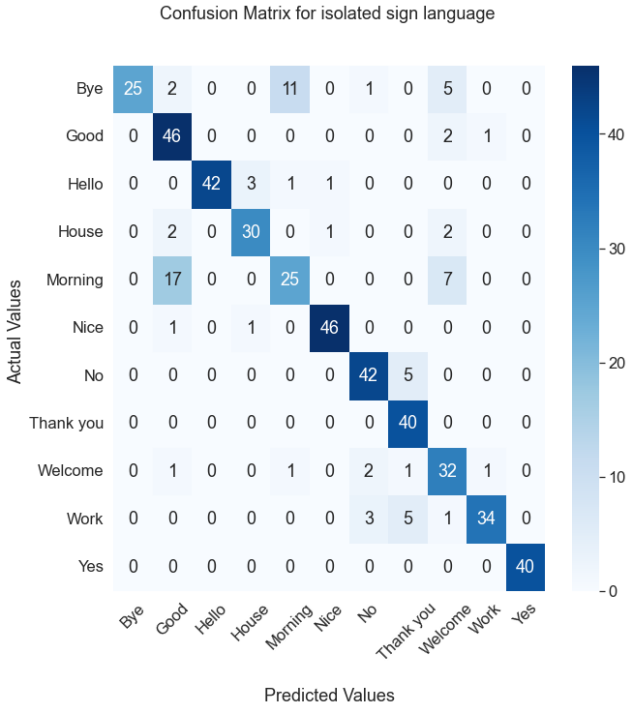

FIGURE 12. Comparative analysis of Proposed Architecture with benchmark dataset of isolated sign language.

model. Simulation of proposed methodology on validation data illustrated as confusion matrix for 11 Indian Sign gestures in figure 12. Apart from the Indian isolated sign dataset, proposed architecture finds remarkable performance on other benchmark isolated datasets like AUTSL and GSL as illustrated in figure 13 with the roc curve.

V. CONCLUSION

The proposed deep learning model helps to reduce the communication barrier for impaired communities by identifying sign gestures. The proposed novel ensemble approach is best suitable for the identification of Indian isolated sign language, which includes 11 distinct signs, based on classification

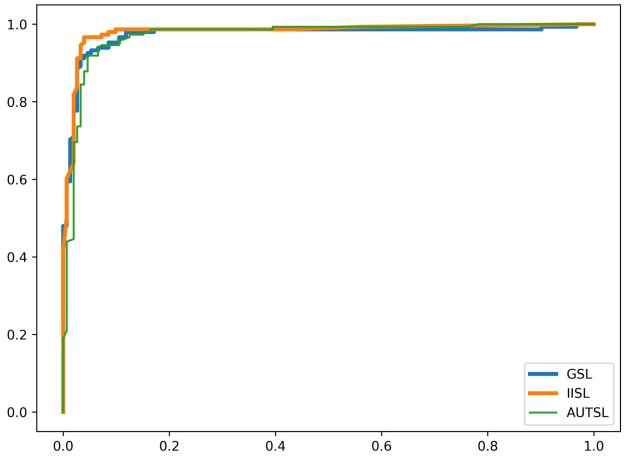

FIGURE 13. Comparative analysis of Proposed Architecture with benchmark dataset of isolated sign language.

efficiency and loss. The state-of-the-art convolutional neural networks and Inception v4 models were standardized and evaluated using ensemble learning techniques. Because it had fewer trainable parameters and a lower computing cost, training the ensemble model appeared simple. As a result, the proposed model seems to be better suited for dynamic sign language recognition, displaying decreased training complexity. The proposed model has achieved a classification accuracy of 98.46%. The proposed architecture was also experimented with a diffident isolated sign language dataset. Ensemble models lead to more complex modes and also large in size. Extension of the proposed study can reduce the size of the model to minimize computational time. This work can be extended to use data augmentation to overcome issues in real-time data collecting and create a multi-object deep learning model that can predict additional sign language classes. Furthermore, an application can be developed with such pre-trained ensemble models.

ACKNOWLEDGMENT

(All the authors contributed equally to this work.)

REFERENCES

[1] V. Wiley and T. Lucas, "Computer vision and image processing: A paper review," *Int. J. Artif. Intell. Res.*, vol. 2, no. 1, p. 22, Jun. 2018, doi: 10.29099/ijair.v2i1.42.

[2] D. Kothadiya, A. Chaudhari, R. Macwan, K. Patel, and C. Bhatt, "The convergence of deep learning and computer vision: Smart city applications and research challenges," in *Proc. 3rd Int. Conf. Integr. Intell. Comput. Commun. Secur.*, Sep. 2021, pp. 14–22. [Online]. Available: https://www.atlantis-press.com/proceedings/iciic-21/125960870

[3] D. R. Kothadiya, C. M. Bhatt, A. Rehman, F. S. Alamri, and T. Saba, "SignExplainer: An explainable AI-enabled framework for sign language recognition with ensemble learning," *IEEE Access*, vol. 11, pp. 47410–47419, 2023, doi: 10.1109/ACCESS.2023.3274851.

[4] M. Hrúz, I. Gruber, J. Kanis, M. Boháček, M. Hlaváč, and Z. Krňoul, "One model is not enough: Ensembles for isolated sign language recognition," *Sensors*, vol. 22, no. 13, p. 5043, Jul. 2022, doi: 10.3390/s22135043.

[5] H. Luqman, "An efficient two-stream network for isolated sign language recognition using accumulative video motion," *IEEE Access*, vol. 10, pp. 93785–93798, 2022, doi: 10.1109/ACCESS.2022.3204110.

[6] E.-V. Pikoulis, A. Bifis, M. Trigka, C. Constantinopoulos, and D. Kosmopoulos, "Context-aware automatic sign language video transcription in psychiatric interviews," *Sensors*, vol. 22, no. 7, p. 2656, Mar. 2022, doi: 10.3390/s22072656.
[7] W. Aditya, T. K. Shih, T. Thaipisutikul, A. S. Fitriajie, M. Gochoo, F. Utaminingrum, and C.-Y. Lin, "Novel spatio-temporal continuous sign language recognition using an attentive multi-feature network," *Sensors*, vol. 22, no. 17, p. 6452, Aug. 2022, doi: 10.3390/s22176452.
[8] K. Kozyra, K. Trzyniec, E. Popardowski, and M. Stachurska, "Application for recognizing sign language gestures based on an artificial neural network," *Sensors*, vol. 22, no. 24, p. 9864, Dec. 2022, doi: 10.3390/s22249864.
[9] M. S. Amin, S. T. H. Rizvi, and M. M. Hossain, "A comparative review on applications of different sensors for sign language recognition," *J. Imag.*, vol. 8, no. 4, p. 98, Apr. 2022, doi: 10.3390/jimaging8040098.
[10] J. J. Bird, A. Ekárt, and D. R. Faria, "British sign language recognition via late fusion of computer vision and leap motion with transfer learning to American sign language," *Sensors*, vol. 20, no. 18, p. 5151, Sep. 2020, doi: 10.3390/s20185151.
[11] Q. Xue, X. Li, D. Wang, and W. Zhang, "Deep forest-based monocular visual sign language recognition," *Appl. Sci.*, vol. 9, no. 9, p. 1945, May 2019, doi: 10.3390/app9091945.
[12] M. S. Amin, S. T. H. Rizvi, A. Mazzei, and L. Anselma, "Assistive data glove for isolated static postures recognition in American sign language using neural network," *Electronics*, vol. 12, no. 8, p. 1904, Apr. 2023, doi: 10.3390/electronics12081904.
[13] M. Al-Hammadi, M. A. Bencherif, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, W. Abdul, Y. A. Alohali, T. S. Alrayes, H. Mathkour, M. Faisal, M. Algabri, H. Altaheri, T. Alfakih, and H. Ghaleb, "Spatial attention-based 3D graph convolutional neural network for sign language recognition," *Sensors*, vol. 22, no. 12, p. 4558, Jun. 2022, doi: 10.3390/s22124558.
[14] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," *Appl. Sci.*, vol. 9, no. 13, p. 2683, Jul. 2019, doi: 10.3390/app9132683.
[15] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131–153, Aug. 2017, doi: 10.1007/s13042-017-0705-5.
[16] J. Fink, B. Frénay, L. Meurant, and A. Cleve, "LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.
[17] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from RGB video using pose flow and self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3436–3445.
[18] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11303–11312.
[19] M. Boháček and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 182–191.
[20] M. Campos-Taberner, F. J. García-Haro, B. Martínez, E. Izquierdo-Verdiguier, C. Atzberger, G. Camps-Valls, and M. A. Gilabert, "Understanding deep learning in land use classification based on Sentinel-2 time series," *Sci. Rep.*, vol. 10, no. 1, p. 17188, Oct. 2020, doi: 10.1038/s41598-020-74215-5.
[21] A. Armagan et al., "Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3D hand pose estimation under hand-object interaction," in *Proc. 16th Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 12368, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 85–101, doi: 10.1007/978-3-030-58592-1_6.
[22] H. Chen, Y. Li, H. Fang, W. Xin, Z. Lu, and Q. Miao, "Multi-scale attention 3D convolutional network for multimodal gesture recognition," *Sensors*, vol. 22, no. 6, p. 2405, Mar. 2022, doi: 10.3390/s22062405.
[23] H. Chen, J. Wang, Z. Guo, J. Li, D. Zhou, B. Wu, C. Guan, G. Chen, and P.-A. Heng, "SignVTCL: Multi-modal continuous sign language recognition enhanced by visual-textual contrastive learning," 2024, *arXiv:2401.11847*. Accessed: Jun. 24, 2024.
[24] K. M. H. H. Rawf, A. O. Abdulrahman, and A. A. Mohammed, "Improved recognition of Kurdish sign language using modified CNN," *Computers*, vol. 13, no. 2, p. 37, Jan. 2024, doi: 10.3390/computers13020037.
[25] L. Al Khuzayem, S. Shafi, S. Aljahdali, R. Alkhamesie, and O. Alzamzami, "Efhamni: A Deep learning-based Saudi sign language recognition application," *Sensors*, vol. 24, no. 10, p. 3112, Jan. 2024, doi: 10.3390/s24103112.
[26] V. Singla, S. Bawa, and J. Singh, "Enhancing Indian sign language recognition through data augmentation and visual transformer," *Neural Comput. Appl.*, pp. 1–14, May 2024, doi: 10.1007/s00521-024-09845-1.
[27] J. M. Joshi and D. U. Patel, "GIDSL: Indian–Gujarati isolated dynamic sign language recognition using deep learning," *Social Netw. Comput. Sci.*, vol. 5, no. 5, pp. 1–12, May 2024, doi: 10.1007/s42979-024-02776-7.
[28] S. Dhulipala, F. F. Adedoyin, and A. Bruno, "Sign and human action detection using deep learning," *J. Imag.*, vol. 8, no. 7, p. 192, Jul. 2022, doi: 10.3390/jimaging8070192.
[29] A. Jana and S. S. Krishnakumar, "Sign language gesture recognition with convolutional-type features on ensemble classifiers and hybrid artificial neural network," *Appl. Sci.*, vol. 12, no. 14, p. 7303, Jul. 2022, doi: 10.3390/app12147303.
[30] J. R. Alvarez, M. Arroqui, P. Mangudo, J. Toloza, D. Jatip, J. M. Rodriguez, A. Teyseyre, C. Sanz, A. Zunino, C. Machado, and C. Mateos, "Estimating body condition score in dairy cows from depth images using convolutional neural networks, transfer learning and model ensembling techniques," *Agronomy*, vol. 9, no. 2, p. 90, Feb. 2019, doi: 10.3390/agronomy9020090.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[32] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sens.*, vol. 11, no. 9, p. 1068, May 2019, doi: 10.3390/rs11091068.
[33] J. Cao, M. Yan, Y. Jia, X. Tian, and Z. Zhang, "Application of a modified inception-v3 model in the dynasty-based classification of ancient murals," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, pp. 1–25, Jul. 2021, doi: 10.1186/s13634-021-00740-8.
[34] L. Alzubaidi et al., "A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications," *J. Big Data*, vol. 10, no. 1, p. 46, Apr. 2023, doi: 10.1186/s40537-023-00727-2.
[35] W. Song and G. Zhang, "Risky-Driving-Image recognition based on visual attention mechanism and deep learning," *Sensors*, vol. 22, no. 15, p. 5868, Aug. 2022, doi: 10.3390/s22155868.
[36] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.-B. Gil-González, and J. M. Corchado, "Deepsign: Sign language detection and recognition using deep learning," *Electronics*, vol. 11, no. 11, p. 1780, Jun. 2022, doi: 10.3390/electronics11111780.
[37] D. R. Kothadiya, C. M. Bhatt, T. Saba, A. Rehman, and S. A. Bahaj, "SIGNFORMER: DeepVision transformer for sign language recognition," *IEEE Access*, vol. 11, pp. 4730–4739, 2023, doi: 10.1109/ACCESS.2022.3231130.
[38] O. M. Sincan and H. Y. Keles, "AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181340–181355, 2020, doi: 10.1109/ACCESS.2020.3028072.
[39] L. Meng and R. Li, "An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network," *Sensors*, vol. 21, no. 4, p. 1120, Feb. 2021, doi: 10.3390/s21041120.
[40] Z. Liang, H. Li, and J. Chai, "Sign language translation: A survey of approaches and techniques," *Electronics*, vol. 12, no. 12, p. 2678, Jun. 2023, doi: 10.3390/electronics12122678.
[41] A. C. Caliwag, H.-J. Hwang, S.-H. Kim, and W. Lim, "Movement-in-a-video detection scheme for sign language gesture recognition using neural network," *Appl. Sci.*, vol. 12, no. 20, p. 10542, Oct. 2022, doi: 10.3390/app122010542.
[42] W. Choi and S. Heo, "Deep learning approaches to automated video classification of upper limb tension test," *Healthcare*, vol. 9, no. 11, p. 1579, Nov. 2021, doi: 10.3390/healthcare9111579.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[44] B. Aksoy, O. K. M. Salman, and Ö. Ekrem, "Detection of Turkish sign language using deep learning and image processing methods," *Appl. Artif. Intell.*, vol. 35, no. 12, pp. 952–981, Sep. 2021, doi: 10.1080/08839514.2021.1982184.
[45] J. Shin, A. S. M. Miah, M. A. M. Hasan, K. Hirooka, K. Suzuki, H.-S. Lee, and S.-W. Jang, "Korean sign language recognition using transformer-based deep neural network," *Appl. Sci.*, vol. 13, no. 5, p. 3029, Feb. 2023, doi: 10.3390/app13053029.
[46] D. Kothadiya, C. Bhatt, D. Soni, K. Gadhe, S. Patel, A. Bruno, and P. L. Mazzeo, "Enhancing fingerprint liveness detection accuracy using deep learning: A comprehensive study and novel approach," *J. Imag.*, vol. 9, no. 8, p. 158, Aug. 2023, doi: 10.3390/jimaging9080158.

DEEP R. KOTHADIYA received the bachelor's and master's degrees in computer science and engineering from Gujarat Technological University and the Ph.D. degree from the Charotar University of Science and Technology (CHARUSAT). He is currently an Assistant Professor with the U & P U Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Technology, CHARUSAT. He is also a Research Scholar with Prince Sultan University, Riyadh, Saudi Arabia. He has already published many research articles in SCI-indexed journals. He is also a Technical Reviewer of many SCI and Scopus-indexed journals, including IEEE Access, *Journal of Visual Communication and Image Representation*, and *Scientific Reports*. He has also been a reviewer of many international conferences, including IEEE and Springer.

CHINTAN M. BHATT (Senior Member, IEEE) was an Assistant Professor with the CE Department, CSPIT, CHARUSAT, for 11 years. He is currently an Assistant Professor with the Department of Computer Science and Engineering (CSE), School of Technology, Pandit Deendayal Energy University (PDEU). He was involved in the successful organization of a few special issues in SCI/Scopus journals. He is the author or co-author of more than 80 publications in the areas of computer vision, the Internet of Things, and fog computing. He has won several awards, including the CSI Award and the Best Paper Award for his CSI articles and conference publications.

HENA KHARWA is currently pursuing the bachelor's degree in computer engineering with the Charotar University of Science and Technology (CHARUSAT). She is a Student with the U & P U Patel Department of Computer Engineering, where she actively engages in both theoretical learning and practical application. Her academic endeavors have led her to contribute significantly to various projects in the fields of deep learning and machine learning. These projects, submitted to the esteemed Chandubhai S. Patel Institute of Science and Technology, serve as tangible evidence of her intellectual curiosity and innovative approach to problem-solving.

FELIX ALBU (Senior Member, IEEE) received the B.Sc. degree in electronics and the Ph.D. degree in telecommunications from the Politehnica University of Bucharest, in 1993 and 1999, respectively, and the Dr.Habil. degree in electronics and telecommunications, in 2014. He was a Teaching Assistant with the Politehnica University of Bucharest, from 1993 to 1999. During the Ph.D. studies, he was a Visiting Researcher for about two years with the National Institute of Telecommunications, Evry, France, and LAAS-CNRS, Toulouse, France. He has obtained an extensive research experience as a Postdoctoral Researcher with University College Dublin, Ireland, from 1999 to 2002; and the Aristotle University of Thessaloniki, Greece, from 2004 to 2005. He got industrial research experience with Lake Communications, Dublin, from 2002 to 2003, and Fotonation Romania, from 2006 to 2011. He is currently a Professor with the Valahia University of Târgovişte, Romania. His Hirsch index is 31 according to Google Scholar, 25 according to Scopus, and 20 according to Web of Science. Since 2013, he has been an IEEE SPS Senior Member. He has been an Associate Editor of *Pattern Analysis and Application* (Springer), since 2015, *Shock and Vibration* (Wiley), since 2018, IEEE Access, *PLOS One*, and *Journal of Intelligent & Fuzzy Systems*, since 2021.