International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

ISL-CNN: A CNN based Automated System for the Recognition of Indian Sign Language for Hearing-Impaired

Reshna S[1], Haris P A[2], Imthias Ahamad T P[3], Jayaraju M[4]

[1]TKM College of Engineering, Kollam, Kerala, India
Corresponding Author Email: *reshna.s[at]tkmce.ac.in*

[2]College of Engineering, Thiruvananthapuram

[3]TKM College of Engineering, Kollam, Kerala, India

[4]UKF College of Engineering and Technology, Paripally, Kollam, Kerala, India

Abstract: ***Sign language is a visual/ gestural language used by people with hearing disabilities. It uses specific shapes and movements of the hands, arms and fingers along with movements of the head, face and eyes. Sign Language Recognition System is an automated system that can translate sign language into spoken language or text. Indian Sign Language (ISL) uses both hands to make gestures to represent most of the signs and one hand moves faster than the other at times in dynamic hand gestures. It involves both global and local hand motions. To determine all these aspects, the position of hands and head, configuration (angles and rotations), and movement (velocities) need to be identified. In this study, we developed ISL_CNN architecture to interpret signs in ISL. We used the dataset developed by Robotics and AI Lab, IIIT, Allahabad. We implemented a modified version of VGG16 Convolutional Neural Networks (ISL _CNN) for the classification of ISL signs of the English alphabet and isolated signs of 23 words. The dataset contained images of both static and dynamic signs. In our study, we used 11 dynamic signs and 9 static signs. The accuracy obtained for the alphabet datasets was 99.81% with 0.0034 loss and that of the ISL words dataset was 99.48% with 0.021 loss. The proposed system may be improved to predict all signs in the ISL dictionary by adding new words and terms, thus making the hearing-impaired person more independent. Additionally, a text-to-speech engine can convert these predicted words into speech.***

Keywords: Convolutional Neural Networks, Indian Sign Language Recognition System, ISL_CNN, Static and Dynamic Gestures, VGG16

1. Introduction

Sign language is a gestural/visual language for vocally challenged people that use specific shapes and movements of the fingers, hands, arms, and movements of the eyes, face, head, and body. Individuals hesitate to learn sign language to communicate with deaf and hard-of-hearing people. A translator is preferred when conversing with the hearing-challenged, but such qualified and trained interpreters are limited. A device that can interpret sign language into spoken language might help the hearing-challenged to interact with others. No internationally recognized and standardized sign language exists for all deaf and hard-of-hearing people. As in spoken language, every country has its sign language with many grammatical variations. Indian Sign Language (ISL) is the language commonly practiced by the hearing-impaired community of India. The Ministry of Social Justice & Empowerment has launched the India Sign Language Dictionary, which was developed by the Indian Sign Language Research & Training Centre (ISLR&TC) under the Department of Empowerment of Persons with Disabilities (DEPwD) (https://www. islets.nic.in/). Recognition of sign language is significant both technically and in terms of its impact on society.

The ISL dictionary contains different categories of words – legal, medical, academic, technical, and daily use words. Most of the words in ISL are dynamic in nature, involving the movement of the hands and head. Facial expressions play a vital role in sign language communication which differentiates the various moods of situation. The sign may contain beats, deictic gestures, iconic gestures and effects on the facial expressions. Beats are rhythmic and often repeating flicks (short and rapid) of the hand or the fingers as in the sign for morning; the closed hand is going upward and opening to symbolise 'morning'. 'Happy' is represented by similar hand movements. Deictic gestures are pointing motions that can be physical (pointing to a location, object, or person) or abstract (identifying an abstract location, period) as the sign for 'you', 'there', and so on. Iconic gestures are hand movements that indicate a figural description or an activity (for example, a hand going either upwards or downwards with wiggling fingers to signify climbing upwards or below). A sign may contain effects with facial expressions that indicate the imparted emotion or communicator's intentions (for example, to symbolise happy/sad and very happy/sad, the same gesture for happy/sad with different facial expressions is used to convey the emotion).

The main features of ISL are
1) ISL uses both hands to make gestures that represent most of the alphabet.
2) ISL uses static and dynamic hand gestures
3) Facial expressions are also included.
4) One hand moves faster than the other hand at times during dynamic hand gestures.
5) Many of the gestures result in obstruction.
6) Complicated hand shapes.

7) The locations of the hand with respect to the body contribute to the Sign.
8) Head/body postures.
9) ISL Involves both global and local hand motion.

Sign language makes use of gestures and body movements simultaneously in the spatial and temporal spaces, which can be identified as both static and dynamic. In static gestures, the location of the hands and fingers in space is fixed, with no movement relative to time, whereas dynamic gestures have a continuous movement of the hands relative to time. Sign language recognition can be classified based on body gestures (manual or non-manual), acquisition methods (vision, sensor, depth, or hybrid), sign category (fingerspelling, isolated, or continuous), detection methods (static/non-tracking, dynamic/tracking, or segmentation), recognition techniques (machine learning or non-machine learning), and classification conditions (signer-dependent or signer independent). Researchers have used machine language techniques such as Support Vector Machines (SVM), K nearest neighbor (KNN), Hidden Markov Models (HMM), Artificial Neural Networks (ANN) and ensembled methods for the recognition of sign languages. Recently, deep neural network architectures, such as CNN, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) networks, have been used for hand gesture recognition.

Hand gesture recognition for sign language conversion can be classified into sensor-based or vision-based methods. In the sensor-based process, the signer has to wear an electronic sensor-based circuitry such as data gloves, accelerometer, and band. These will measure the movement of the hands and send the particulars to the computer for further processing. This approach has shown good results in the literature, but it is inconvenient for the users and expensive. The vision-based method captures an image of the signer using a camera. This method uses image processing algorithms to process the captured image and reduces the dependency on sensory devices. This paper proposes a vision-based method using a modified VGG16 convolutional neural network (ISL_CNN) to recognize static and dynamic hand gestures in Indian sign language. The proposed CNN model was found to be more beneficial than the present state-of-the-art methods because it has high accuracy and consumes less training time.

The major contribution and novelty of the paper are as follows:

- For the recognition of Indian Sign Language, a solid model has been put forth that includes both static and dynamic signs. ISL signs are made with complex hand forms and are made with both hands.
- The study made use of a dataset that included 23 ISL words and 26 English alphabets. The dataset was developed at Robotics and AI Lab, Indian Institute of Information Technology (IIIT), Allahabad. It was created in an environment with a consistent background and different lighting levels.
- The hyper-parameters of the model, including kernel width, epochs, batch size, and learning rate, are all changed analytically to ensure effective training.
- Several evaluation metrics, including accuracy, loss, the recognition accuracy of each class, and model training time consumption were used to conduct a thorough experimental assessment of the suggested work.
- The augmented dataset was used to assess the performance of the model. Given that the model is invariant to rotation and scaling changes, it produces competent findings and is considered to be resilient.

The rest of the paper is organized as follows: Section II discusses the different deep learning techniques in the literature. Section III describes the features of the ISL, the dataset used, and the proposed methodology. Section IV focuses on the experimental outcomes and Section V presents the conclusions.

2. Related Works

Deep neural network architectures, such as CNN and the long short-term memory (LSTM) network, have recently been used for hand gesture recognition. Ahmed KASAPBAS et al. developed a dataset of the American Sign Language alphabet (ASLA) and a Convolutional Neural Network-based sign language interface system to interpret gestures of sign language and hand poses in natural language [1]. Various conditions, such as lighting, distance, and other non-variable conditions, were considered when developing the dataset. They tested the model with three different American Sign Language alphabet datasets and achieved 99.38% accuracy with excellent prediction and a slight loss (0.0250).

Convolutional neural networks and machine learning algorithms were used by Sharma A., et al. to compile a thorough comparative analysis of several gesture detection approaches and assess their real-time accuracy [2]. Based on several trainable parameters, three models—a pre-trained VGG16 with fine-tuning, a VGG16 with transfer learning, and a hierarchical neural network—were examined. These models were trained using a self-created dataset comprising pictures of the ISL renditions of each of the 26 letters of the English alphabet. The performance evaluation was simulated by varying the lighting and background environments. The hierarchical model fared better than the other two models of the three, with the best accuracy of 98.52% for one-hand gestures and 97% for two hand gestures [3]. This model was used to create a chat interface in Django that converts gestures into voice and vice versa in real time.

Garcia et al. proposed a real-time ASL recognition system that uses Convolutional Neural Networks (CNN) to translate a video of a user's ASL signs into text [4]. They used a pre-trained GoogleNet CNN. Their model correctly classified letters A – K. A real-time American sign language (ASL) recognition system was developed using a pre-trained AlexNet and VGG CNN and tested by Sahoo J.P. et al. [5]. The effectiveness of the proposed technique was evaluated using leave-one-subject-out cross-validation (LOO CV) and regular cross-validation CV tests on the Massey University (MU) dataset and HUST American Sign Language (HUST-ASL) datasets. Mean accuracies of 98.14% and 64.55% were obtained for both datasets. Yirtici, Tolga, and Kamil Yurtkan proposed a regional-CNN-based technique to recognize Turkish sign language [6]. They have used a pre-trained

AlexNet network. The new model was trained using a region-based Convolutional Neural Network (R-CNN) object detector. The system achieves 99.7% an average precision.

K. H. Rawf et al. suggested a real-time model for recognising Kurdish sign language alphabets using a CNN algorithm [7]. The model was trained and forecasted on the KuSL2022 dataset over a number of epochs using various activation functions. The dataset contains 71,400 images from two separate datasets for the 34 Kurdish sign languages and alphabets. The results reveal that the suggested system improved its classification and prediction model performance, with an average training accuracy of 99.91%. Abdul Mannan et al. proposed a robust ASL recognition system that involved the signs of 24 alphabets [8]. The proposed method is based on deep convolutional neural networks that can recognize the ASL alphabets with an accuracy of 99.67% on unseen test data. Batool Yahya AlKhuraym et al. adapted Efficient Network (EfficientNet) models to classify Arabic Sign Language gestures [9]. The dataset was developed using different signers with background variations for thirty different Arabic alphabets. They achieved 94% accuracy using the EfficientNet-Lite 0 architecture and the Label Smooth as the loss function.

From the literature review presented, the following interpretations are observed:

- In contrast to other widely used sign languages, the ISL has a more complex format for gestures. Therefore, applying an existing gesture recognition system to ISL will not produce the same outcomes.
- The gesture recognition method for the recognition of ISL has drawn significantly less attention because of the complex structure of ISL.
- The ISL dataset used in sign language recognition systems in the literature was collected by the respective authors only. They are limited in size and are acquired under limited background conditions and signers.
- The process of converting the crucial information in the input data into a small feature vector is known as feature extraction. Traditional feature extraction methods, which are used with machine learning models, require mathematical operators and manual observation key feature extraction. Examples include the Shift invariant feature transform (SIFT), Principal component analysis (PCA), Histogram of Oriented Gradient (HOG), Local binary pattern (LBP), etc. These calculations are complex in nature. For a few ISL classes, the accuracy reported in the literature is insufficient. In contrast, deep learning automates feature extraction. The model automatically learns and extracts the relevant properties from the input data with each additional layer of the neural networks. This automatic feature extraction using deep learning has an advantage over the feature extraction algorithms.

3. Methodology

a) Dataset

In this study, we used a custom dataset developed by Robotia Lab, Indian Institute of Information Technology, Allahabad (https://robita.iiita.ac.in/). It contains images of 23 words and 26 English alphabets of the ISL signs. The dataset was created using an external camera Canon EOS (single camera) with an 18–55 mm lens, 18 megapixels, 29 frames per second, and a resolution of 3920*320 bits/sec. It was developed under different light illumination contrasts with background uniformity; a dark background was chosen to effectively handle grayscale images [10]. They have considered only the upper body parts when developing the dataset. They extract the foreground image from the complete image by removing the background to obtain the silhouette of the upper body part. The hand region is then subtracted from these foreground images by eliminating the face to obtain the hand portion from the upper body.

Initially, each video is converted into a sequence of RGB frames. Each frame has dimensions of 640 x 480. Skin colour segmentation was applied to extract the skin region. To find the skin region, each frame is converted into the HSV (Hue, saturation, value) plane, where only H and S values with a threshold ($H > 0.55$ or $S <= 0.20$ or $S > 0.95$) were used for finding the non-skin region of an image and is eliminated to attain the skin region. A median filter was applied to preserve the edges of the segmented region. It mainly removes salt pepper noise and impulsive noise for edge preservation. Images obtained after median filtering were converted into binary form. This process was followed by a histogram equalization technique for normalization. Sample images of the alphabet are shown in Fig.1

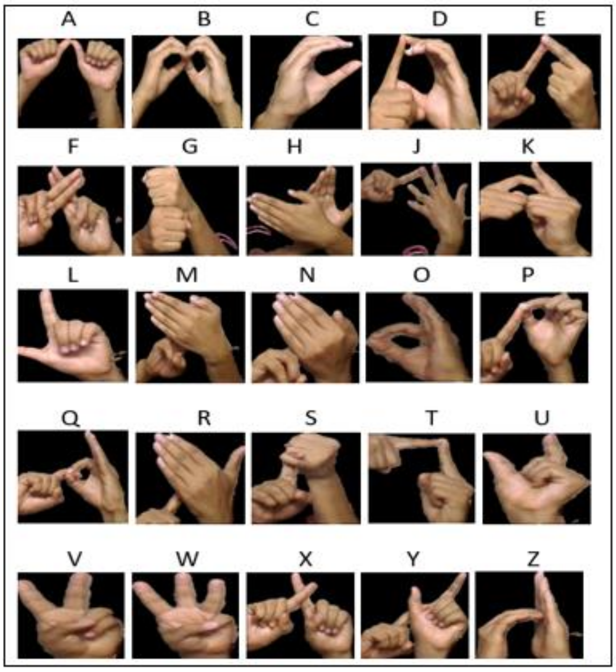

Figure 1: Alphabets in ISL

The dataset consists of both static and dynamic hand gestures. The alphabet signs are static. There are 400 images of each alphabet signs, thus there are a total of 10426 images in the alphabet dataset. Out of 23 signs of words in the dataset, 12 were static and 11 were dynamic in nature. "ABROAD", "ASCEND", “ALL-GONE”, "BESIDE", “DRINK”, "FLAG", "HANG", "MARRY", "MIDDLE", "MOON", and "PRISONER" are static gestures and the signs for "ABOVE", "ACROSS", "ADVANCE", "AFRAID", "ALL", "ALONE", "ARISE", "BAG", "BELOW", "BRING", and "YES" are dynamic gestures. Static images are shown in Fig. 2 and the dynamic gesture images are shown in Fig. 3.

The dataset of dynamic images includes the different hand shapes of the sign from the start frame to the end frame of the video. There are more than 1500 images of each dynamic signs and nearly 1000 images of each static signs, thus there are a total of 38215 images in the gesture dataset. We have trained and tested the two datasets separately with 70% of the images for training, 15% for testing and 15% for validation. To the best of our knowledge, there is currently no study has reported on the recognition of this dataset using Convolutional Neural Networks. Since each dataset in the field of deep learning has unique features that can be used to enhance existing models, the creation of a new CNN may be considered as a fresh addition to the field.

Figure 2: Static Gestures

b) System Architecture

The capacity of CNNs to recognize patterns and make sense of them has significantly changed how they approach picture recognition. They are considered to be the most efficient architectures for image classification, retrieval, and detection tasks due to the high level of accuracy in their outputs. The ability of CNNs to achieve "spatial invariance," which means they can learn to identify and extract visual information from any point in the image, is a significant capability. There is no need for separate feature extraction because CNNs automatically learn characteristics from the images/data and perform image extraction. CNNs are a powerful deep learning methods for generating accurate results.

Convolutional Neural Networks (CNNs) are a type of deep learning algorithms that are particularly adept at processing and identifying images. Convolutional, pooling, and fully connected layers, among others, constitute its structure. The crucial component of a CNN is its convolutional layers, where filters are used to extract features from the input image such edges, textures, and shapes. The Convolutional layer output was then passed via pooling layers, which were used to down-sample the feature maps and retain the most important information while reducing the spatial dimensions. The output of the pooling layers was then applied to one or more fully connected layers to forecast or categorize the image. The architecture of the CNN is shown in Fig.4. CNNs are trained to recognize the patterns and characteristics that are associated with specific objects or classes using a large collection of labelled images. In addition to being used to extract features for other tasks such as object detection or image segmentation, a CNN may be taught to categorize new images.

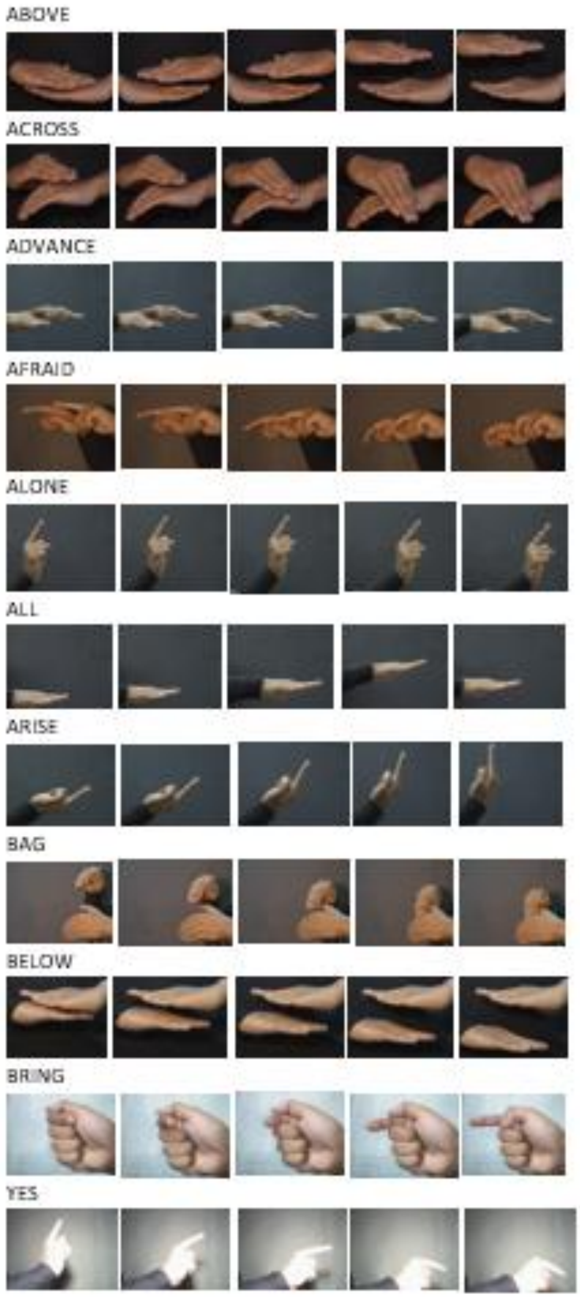

Figure 3: Dynamic Gestures

In our experiments, we have used a modified version of the VGG16 CNN to implement the Indian Sign Language Recognition System. VGG-16 is one of the biggest networks with 138 million parameters [9]. Default VGG-16 accepts colored images of dimensions 227×227 and outputs with 1000 classes. In our system, we have modified the weights of the pre-trained VGG 16 to classify 26 English Alphabets and 23 ISL words.

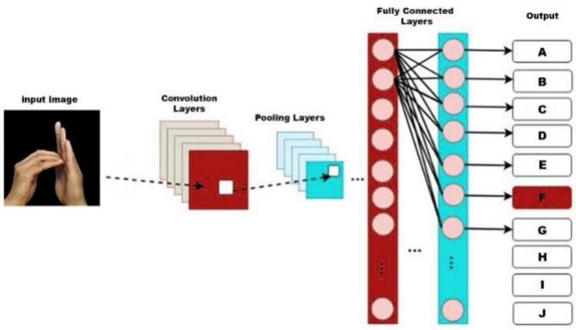

Figure 4: CNN Architecture

The proposed ISL_CNN has 12 layers: 10 convolution layers with a kernel size 3×3 and three fully connected layers. The architecture of the proposed ISL_CNN is shown in Fig.5. A convolution layer is a means that allows feature extraction of images based on several filters trained by the network itself. Moreover, an activation function is applied to allow the network learn more complex patterns. ReLu (Rectified Linear Unit) is the most frequently used activation function. The output is zero for all inputs less than or equal to zero and is the same as the input for inputs greater than zero. The equation for ReLU is given in Equation (1).

$$f(x) = \max(0, x) \tag{1}$$

All the convolution layers use ReLU as their activation function. This results in faster learning and decreases the likelihood of the vanishing gradient problems.

The first two layers are convolution layers with 64 channels of a 3×3 filter size and the 'same' padding (padding of 1 pixel). The next is a max-pooling layer of stride (2, 2). The 4th and 5th layers are convolution layers with 128 filter size and filter size (3, 3). This is followed by a max-pooling layer of stride (2, 2) which is the same as in the previous max-pooling layer. Next, there are three convolution layers of filter size (3, 3) and 256 filters, followed by a max-pooling layer of stride (2, 2). Then, there are three convolution layers of filter size (3, 3) and 512 filters and a max pool layer with the same padding. After the sets of convolution and max-pooling layer, there was a (7, 7, 512) feature map. This output was flattened to make it a (1, 25088) feature vector. After this there are three fully connected dense layers, the first layer takes input from the last feature vector and outputs a (1, 4096) vector, the second layer also outputs a vector of size (1, 4096) and 3rd fully connected layer is used to implement the softmax function to classify 26 classes. The Softmax activation function calculates the relative probabilities for each class. The equation for softmax activation function is given in Equation 2.

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{c}(e^{z_j})} \tag{2}$$

where z represents the values from the neurons of the output layer and c is the number of classes. The exponential function acts as a non-linear function. It is normalized by dividing the values by the sum of the exponentials and then converting them into probabilities.

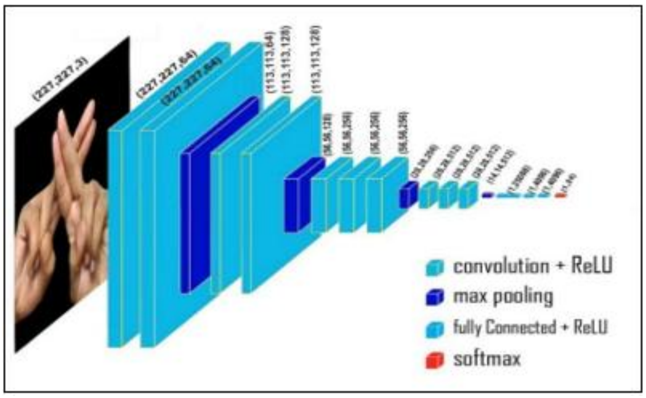

Figure 5: Architecture of the proposed ISL_CNN

The loss function used for modelling is categorical cross-entropy. A loss function measures how well our predicted class labels match our ground-truth labels. The greater the degree of agreement among those sets of labels, the lower the loss. The activation function employed in the output layer of the neural network is directly tied to the preference for the loss function. Categorical cross-entropy is a loss function that is used in multiclass classification applications. These are tasks in which an instance can only belong to one of many possible classes, and the model must determine which one. Every predicted option is compared with the actual output value (0 or 1), and a score is computed to penalizes the probability based on the difference from the predicted value. The equation for categorical cross entropy is given by Equation (3)

$$L_{CE} = \sum_i T_i \log(S_i) \tag{3}$$

where T_i is the true value with values 0 and 1 and S_i is the softmax probability for the i^{th} class. The softmax function is continuously differentiable. This enables the derivative of the loss function to be calculated for each weight in the neural network. Owing to this property, the model can modify the weights in a way that minimises the loss function and produces results that are close to the true values. On each iteration, the model parameters were updated, the loss function was reviewed and the version was updated after each training sample. These frequent updates result in faster convergence to minima, but at the expense of amplified variance, which may cause the model to exceed the desired location. The optimizer used was Adaptive Moment Estimation (ADAM), which uses estimations of the first and second moments of the gradient to adapt the learning rate for each weight of the neural network. Adam was proposed as the most efficient stochastic optimization which only requires first order gradients where the memory requirement is too small. In addition, in Adam, the hyperparameters have instinctive interpretations and hence required less tuning [12]. Adam performs well.

4. Results and Discussion

The proposed method was implemented with ISL datasets for 26 English alphabets and 20 signs taken from the ISL Dictionary. The experiments were run with a 64-bit 11th Gen Intel(R) Core (TM) i5-1135G7, 2.40GHz with 1.38 GHz RAM on the Google Colab platform. We implemented our system by using PYTHON and OpenCV. We have developed two different CNNs to classify 26 English Alphabets and the ISL signs of 23 words. The performance of the proposed ISL_CNN was also evaluated using augmented dataset.

This was performed to make the trained models more applicable. Data augmentation is a method for generating fresh samples from the datasets. In this study, rotation and scaling techniques were used to produce four additional samples for each signer sample. For this, a random rotation between the [-20] and [+-20] as well as a random scaling of [0.8-1.5] inside and outward were applied.

To evaluate the models, commonly used measures such as accuracy, precision, recall, and confusion metrics are considered.

The accuracy is the ratio of the number of correct predictions to the total number of predictions made for a dataset. It is a valid choice of evaluation for classification problems that are well balanced and not skewed or have no class imbalance.

$$\text{Accuracy} = \frac{\text{Correct prediction}}{\text{Total prediction}}$$
$$= \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

Precision refers to the number of correctly predicted cases that turned out to be positive. It is useful for the skewed and unbalanced datasets. The higher the number False positives predicted by the model, the flower the precision.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall refers to the number of actual positive cases that could be predicted correctly with our model. This is also known as sensitivity or hit rate. This measures the ability of the model to detect positive samples. The more false negatives the model predicts, the lower the recall.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

F1-score represents the harmonic mean of the recall and precision. The value of the F1-score ranges from zero to one. A high score indicates that our model generalizes well and exhibits a good performance. This metric only favors classifiers with similar precision and recall.

$$\text{F1-score} = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

The Confusion Matrix is a is a table with combinations of predicted and actual values. This is a n×n matrix, where n is the number of classes. It is often used to describe the performance of a classification model on a set of test data for which the true values are known.

a) CNN for recognition of ISL Signs of English Alphabets

A modified version of VGG 16 was used to classify the alphabet signs. We have trained the model on our dataset and new weight values are generated. The dataset is split in 70:15:15 ratio for training, testing, and validation respectively. There are 400 images of each alphabet sign, thus there are a total of 10426 images in the alphabet dataset, out of which 7280 images are for training, 1560 images for testing, and 1586 images for validation. The model developed for Alphabet recognition is shown in Table. 1

Table 1: Structure of CNN model developed for Alphabet recognition

Layer type	Output shape	Parameters
input_1 (InputLayer)	(None, 227, 227, 3)	0
block1_conv1 (Conv2D)	(None, 227, 227, 64)	1792
block1_conv2 (Conv2D)	(None, 227, 227, 64)	36928
block1_pool (MaxPooling2D)	(None, 113, 113, 64)	0
block2_conv1 (Conv2D)	(None, 113, 113, 128)	73856
block2_conv2 (Conv2D)	(None, 113, 113, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense	(None, 26)	106522
Total params:		13,43,67,066
Trainable params:		13,43,67,066
Non-trainable params:		0

The program was run for 10 epochs with a batch size of 64. Accuracy, loss, Precision, F1 score, support and Confusion Matrix are calculated for each alphabet. The accuracy attained for the alphabet datasets was 99.81% with 0.0034 loss. The Accuracy plots and loss plots are shown in Fig. 6(a) and 6(b) respectively.

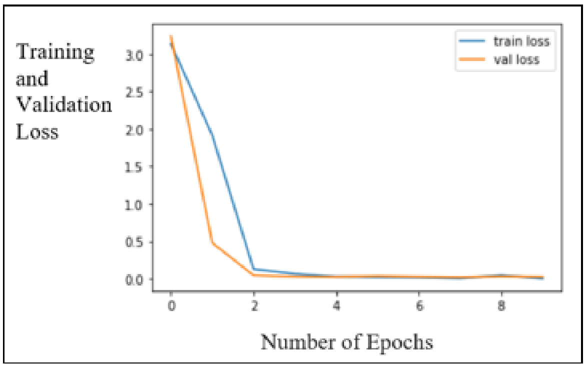

Figure 6 (a): Training and Validation Loss plot

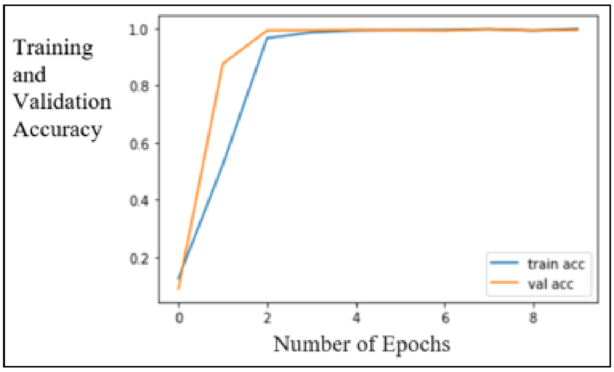

Figure 6 (b): Training and Validation Accuracy plot

The model is evaluated by calculating the precision, recall and F1 score for each class of alphabets and is given below in Table 2.

Table 2: Evaluation metrices of CNN model for the recognition of ISL signs of English alphabets

Class	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	61
1	1.00	1.00	1.00	61
2	1.00	1.00	1.00	61
3	1.00	1.00	1.00	61
4	0.98	1.00	0.99	60
5	0.98	1.00	0.99	60
6	1.00	1.00	1.00	61
7	1.00	1.00	1.00	61
8	1.00	1.00	1.00	61
9	0.98	1.00	0.99	60
10	1.00	1.00	1.00	61
11	1.00	1.00	1.00	61
12	1.00	1.00	1.00	61
13	1.00	1.00	1.00	61
14	1.00	0.97	0.98	63
15	1.00	1.00	1.00	61
16	1.00	1.00	1.00	61
17	1.00	1.00	1.00	61
18	1.00	1.00	1.00	61
19	1.00	1.00	1.00	61
20	1.00	1.00	1.00	61
21	1.00	1.00	1.00	61
22	1.00	1.00	1.00	61
23	1.00	0.98	0.99	62
24	1.00	1.00	1.00	61
25	1.00	1.00	1.00	61
Accuracy			1.00	1586
macro avg	1.00	1.00	1.00	1586
weighted avg	1.00	1.00	1.00	1586

A sample output for the recognition of alphabet A is shown in Fig. 7. The confusion matrix is shown in Fig.8.

b) CNN for recognition of ISL signs of static and dynamic gestures.

We implemented the ISL_CNN model for the recognition of signs of the 23 ISL words. The signs for "ABOVE", "ACROSS", "ADVANCE", "AFRAID", "ALL", "ALONE", "ARISE", "BAG", "BELOW", "BRING", and "YES" are dynamic gestures. Dynamic gestures are represented by short videos. These are divided into different frames from the start to the end of the gesture. There are more than 1500 images in the dataset of dynamic gestures that include all positions of the gestures. "ABROAD", "ASCEND", "ALL-GONE", "BESIDE", "DRINK", "FLAG", "HANG", "MARRY", "MIDDLE", "MOON", and "PRISONER" are static gestures. The dataset contains more than 700 images of these static signs. The model parameters of the ISL_CNN are listed in Table 3.

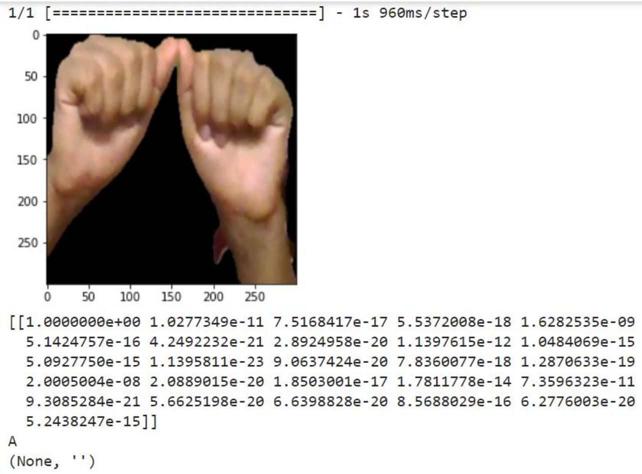

Figure 7: Sample output for the recognition of alphabet A

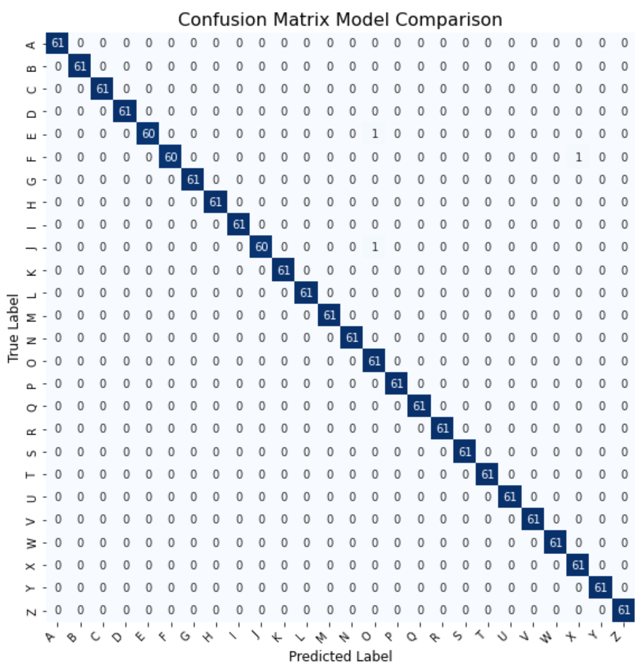

Figure 8: Confusion Matrix for ISL Alphabets

Table 3: Structure of ISL_CNN for the recognition of ISL words

Layer type	Output shape	Parameters
input_1 (InputLayer)	(None, 227, 227, 3)	0
block1_conv1 (Conv2D)	(None, 227, 227, 64)	1792
block1_conv2 (Conv2D)	(None, 227, 227, 64)	36928
block1_pool (MaxPooling2D)	(None, 113, 113, 64)	0
block2_conv1 (Conv2D)	(None, 113, 113, 128)	73856
block2_conv2 (Conv2D)	(None, 113, 113, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080

Paper ID: SR25124140109 DOI: https://dx.doi.org/10.21275/SR25124140109

block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	411045888
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 24)	98328
Total params:		43,55,60,792
Trainable params:		43,55,60,792
Non-trainable params:		0

The program was run for 10 epochs with a batch size of 64. Accuracy, loss, precision, F1 score, support and confusion matrix are calculated for each gesture. The Accuracy plots, and loss plots are presented in Fig. 9(a) and 9(b) respectively. The evaluation metrics are shown in Table 4. The confusion matrix is shown in Fig. 11.

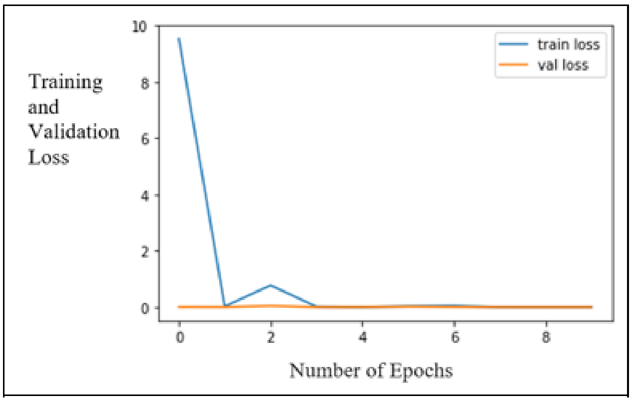

Figure 9 (a): Training and Validation Loss plot

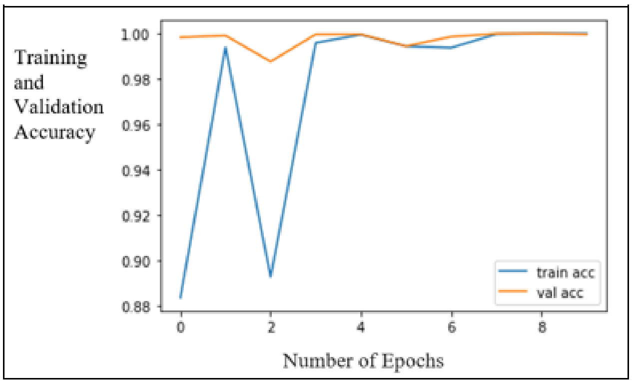

Figure 9 (b): Training and Validation Accuracy plot

Table 4: Evaluation metrices of ISL_CNN model for the recognition of ISL words

Class	Precision	Recall	F1-score	Support
1	1.00	1.00	1.00	216
2	1.00	1.00	1.00	225
3	1.00	1.00	1.00	144
4	1.00	1.00	1.00	147
5	1.00	1.00	1.00	238
6	1.00	1.00	1.00	242
7	1.00	1.00	1.00	121
8	1.00	1.00	1.00	180
9	1.00	1.00	1.00	112
10	1.00	1.00	1.00	227
11	1.00	1.00	1.00	210
12	1.00	1.00	1.00	236
13	1.00	1.00	1.00	241
14	1.00	1.00	1.00	163
15	1.00	1.00	1.00	243
16	1.00	1.00	1.00	188
17	1.00	1.00	1.00	180
18	1.00	1.00	1.00	107
19	1.00	1.00	1.00	74
20	1.00	1.00	1.00	93
21	1.00	1.00	1.00	101
22	1.00	1.00	1.00	85
23	1.00	1.00	1.00	69
accuracy			1.00	3842
macro avg	1.00	1.00	1.00	3842
weighted avg	1.00	1.00	1.00	3842

The predicted output for the ISL word ALL_GONE is shown in Fig. 10.

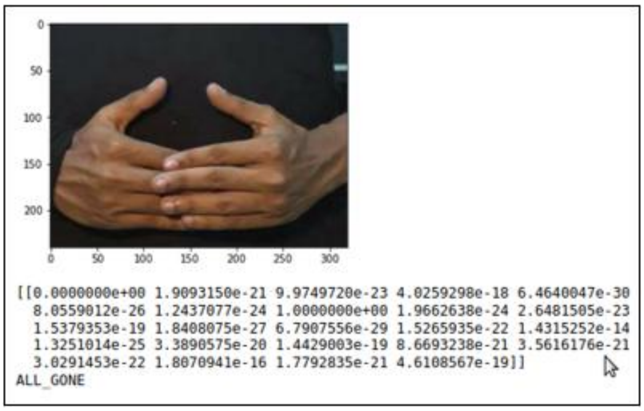

Figure 10: Predicted output for the ISL word ALL_GONE

The performance of the proposed ISL_CNN was compared with that of an existing CNN with the same classification problem of sign language recognition for different sign languages. A comparison was performed based on their achieved accuracy only, as it is the only widely used performance metric among all the state-of-the-art approaches. A comparison is presented in Table. 5. It is evident from these findings that the ISL_CNN model surpasses all the other methods as it achieves the highest accuracy of 99.81% with 0.0034 loss for ISL alphabets and 99.48% with 0.0210 ISL words.

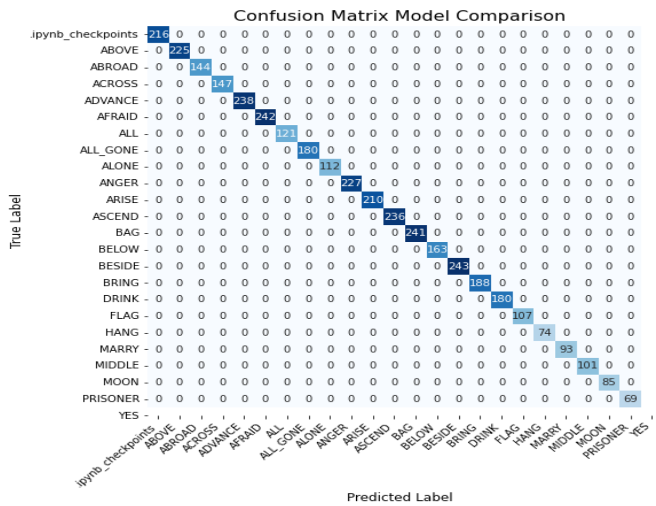

Figure 11: Confusion matrix for ISL words

Paper ID: SR25124140109 DOI: https://dx.doi.org/10.21275/SR25124140109

Table 5: Performance comparison with other CNN models used

Author	Sign Language used	CNN used	Accuracy	Types of gesture
Ahmed KASAPBAS et.al.	American Sign Language alphabet	New model	99.38%	static
Sharma, A., et.al	Indian Sign Language English alphabets	New model G-CNN	97%	static
Garcia, et.al	American Sign Language	pre-trained GoogleNet		static
Sahoo J.P, et.al	Massey University (MU) Dataset and HUST American Sign Language (HUST-ASL) datasets	pre-trained AlexNet and VGG	98.14% 64.55%	
Yirtici, , et.al	Turkish sign language	pre-trained Alexnet	99.7%	
K. H. Rawf et al.	Kurdish sign language alphabets KuSL2022 dataset		99.91%.	
Abdul Mannan, et.al	American Sign Language alphabet		99.67%	
AlKhuraym et.al.	Arabic Sign Language	EfficientNet-Lite 0	94%	
Proposed ISL_CNN	Indian Sign Language English alphabets and words	Modified version of VGG16	99.81%	Static and dynamic

5. Conclusion

In this study, we developed an ISL_CNN architecture to interpret the signs of Indian Sign Language. We developed a CNN for the classification of the ISL signs in 26 English alphabets, 11 dynamic signs, and 12 static signs. The CNN model that we designed provided the best accuracy in the empirical trials. The datasets were developed in the Robotia Lab of the Indian Institute of Information Technology, Allahabad. This dataset may support future research in the field of machine learning and deep learning to develop sign language recognition systems. The accuracy obtained for the alphabet datasets is 99.81% with 0.0034 loss and that of the ISL gesture dataset is 99.48% with 0.0210 In real-world simulation, these accuracies seem competitive, but the most accurate prediction was obtained. As a result, our proposed CNN architecture performs better than earlier SLR models. The addition of images for more words in the collection could enhance this study. To increase accuracy and minimize loss, further images can be inserted. The proposed system can be enhanced to forecast an entire phrase by the inclusion of new words and keywords. Utilizing a text-to-speech engine will also enables the convertion of these predicted words into speech. Future works may be carried out aiming to convert all the signs in the ISL dictionary and to convert a sentence to speech thus making the hearing-impaired person more independent.

Acknowledgement

We would like to thank Robita Lab, Indian Institute of Information Technology, Allahabad, for providing us with the dataset they have developed.

Funding

This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Ahmed KASAPBAS¸ Ahmed Eltayeb AHMED ELBUSHRA, Omar AL-HARDANEE, Arif YILMAZ, "DeepASLR: A CNN based human-computer interface for American Sign Language recognition for hearing-impaired individuals", Computer Methods and Programs in Biomedicine Update, Volume 2, 2022, 100048

[2] Sharma, A., Sharma, N., Saxena, Y. *et al.* Benchmarking deep neural network approaches for Indian Sign Language recognition. *Neural Comput & Applic* **33**, 6685–6696 (2021). https://doi.org/10.1007/s00521-020-05448-8

[3] Garcia, Brandon, and Sigberto Alarcon Viesca. "Real-time American sign language recognition with convolutional neural networks." *Convolutional Neural Networks for Visual Recognition* 2.225-232 (2016): 8.

[4] Sahoo J.P, Prakash A.J, Pławiak P, Samantray S "Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network" Sensors 2022, 22, 706. https://doi.org/10.3390/s22030706

[5] Yirtici, Tolga, and Kamil Yurtkan. "Regional-CNN-based enhanced Turkish sign language recognition." *Signal, Image and Video Processing* (2022): 1-7.

[6] K. H. Rawf, "Effective Kurdish Sign Language Detection and Classification Using Convolutional Neural Networks," 2022, https://doi.org/10.21203/rs.3.rs-1965056/v1.

[7] Abdul Mannan, Ahmed Abbasi, Abdul Rehman Javed, Anam Ahsan, Thippa Reddy Gadekallu, and Qin Xin, "Hypertuned Deep Convolutional Neural Network for Sign Language Recognition", Computational Intelligence and Neuroscience Volume 2022, Article ID 1450822, https://doi.org/10.1155/2022/ 1450822

[8] Batool Yahya AlKhuraym, Mohamed Maher Ben Ismail, Ouiem Bchir, "Arabic Sign Language Recognition using Lightweight CNN-based Architecture", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 13, No. 4, 2022

[9] Simonyan, K. & Zisserman, A., (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

[10] Diederik P. Kingma, Jimmy Lei Ba, Adam: A Method for Stochastic Optimization, International Conference on Learning Representations, ICLR 2015.

[11] S. Sharma and S. Singh, "*Vision-based hand gesture recognition using deep learning for the interpretation of sign language.*" *Expert Systems with Applications*, vol. 182, p. 115657, 2021, doi: 10.1016/j.eswa.2021.115657.

[12] P. D. Hung, N. T. Su. "Unsafe Construction Behavior Classification Using Deep Convolutional Neural Network", Pattern Recognition and Image Analysis, 2021

[13] Batool Yahya AlKhuraym, Mohamed Maher Ben Ismail, Ouiem Bchir. "Arabic Sign Language

Recognition using Lightweight CNNbased Architecture" , International Journal of Advanced Computer Science and Applications, 2022
[14] Kumud Tripathi, Neha Baranwal, G.C. Nandi. "Continuous dynamic Indian Sign Language gesture recognition with invariant backgrounds", 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015
[15] Kumud Tripathi∗, Neha Baranwal and G. C. Nandi, “Continuous Indian Sign Language Gesture Recognition and Sentence Formation”, Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), doi: 10.1016/j.procs.2015.06.060

Author Profile

Prof. Reshna S obtained his B.Tech Degree from Rajiv Gandhi Institute of Technology, Mahatma Gandhi University, Kerala and M.E from P S G College of Technology, Coimbatore, Tamil Nadu. She has 22 years of teaching experience. Presently she is working as an Associate Professor in Department of Electronics & Communication Engineering, TKM College of Engineering, Kollam. Her area of interest are Gesture Identification, Computer Vision and Image Processing

Dr. Haris P A received his B.Tech in Electronics & Communication Engineering from TKM College of Engineering, Kollam, Kerala University and ME in Communication Systems from College of Engineering Guindy. He had obtained his PhD in Wireless Communication from NIT Calicut. He joined as Lecturer in Electronics & Communication Engineering in Directrate of Technical Education, Government of Kerala in 1999. Presently he is working as Professor in ECE at College of Engineering, Trivandum

Dr. Imthias Ahamed T P received the B.Tech. degree in Electrical and Electronics Engineering from Kerala University, India, in 1988 and the M.Tech.(Engg/ Sciences). degree from Calicut University, India, in 1991 and the Ph.D. degree from Indian Institute of Science, Bangalore, India in 2002.From July 1990 to June 2010 and since July 2015, he was a faculty member in the department of Electrical and Electronics Engineering, TKM College of Engineering, Kollam, Kerala, India. He was associated with the Saudi Aramco Chair in Electrical Power, Department of Electrical Engineering, College of Engineering, King Saud University, Riyadh, Saudi Arabia from July 2010 to June 2013 and Department of Electrical and Computer Engineering, College of Engineering, Dhofar University, Salalah, Oman from September 2013 to June 2015.. He is the author of three books and more than thirty articles. His research interests include reinforcement learning applications, power system scheduling, path planning, demand side management, neural networks and deep learning.

Dr. Madhavan Jayaraju received the B.Tech. from the University of Kerala in 1985 and M.Tech. degree in High Voltage Engineering from the Indian Institute of Science in 1994, Bangalore, India and Ph.D. degree from the University of Kerala in 2005. From 1978 to 2012, he was with the Department of Electrical and Electronics Engineering, TKM College of Engineering, Kollam, Kerala, India and later The Director, ANERT, Kerala, India, Principal of M E S Institute of Management and Technology Kollam, Kerala, India, and Principal, College of Engineering, Munnar. He is currently the Dean of the UKF College of Engineering and Technology, Paripally, Kollam, Kerala, India. His research interests include high voltage engineering, energy conservation and management, new and renewable energy sources and computer vision and image processing.