IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 October 2022, accepted 10 December 2022, date of publication 9 January 2023, date of current version 18 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3231130

RESEARCH ARTICLE

SIGNFORMER: DeepVision Transformer for Sign Language Recognition

DEEP R. KOTHADIYA[1], **CHINTAN M. BHATT**[2], **TANZILA SABA**[3], **(Senior Member, IEEE), AMJAD REHMAN**[3], **(Senior Member, IEEE), AND SAEED ALI BAHAJ**[4]

[1]U & P U Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology (FTE), Charotar University of Science and Technology (CHARUSAT), Changa 388421, India
[2]Department of Computer Science and Engineering, School of Engineering and Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat 382007, India
[3]Artificial Intelligence and Data Analytics Lab (AIDA), CCIS, Prince Sultan University, Riyadh 11586, Saudi Arabia
[4]MIS Department, College of Business Administration, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

Corresponding authors: Deep R. Kothadiya (deepkothadiya.ce@charusat.ac.in) and Tanzila Saba (tsaba@psu.edu.sa)

This work was supported by the Artificial Intelligence and Data Analytics Lab (AIDA). College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh, Saudi Arabia. The authors would also like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

ABSTRACT Sign language is the most common form of communication for the hearing impaired. To bridge the communication gap with such impaired people, a normal people should be able to recognize the signs. Therefore, it is necessary to introduce a sign language recognition system to assist such impaired people. This paper proposes the Transformer Encoder as a useful tool for sign language recognition. For the recognition of static Indian signs, the authors have implemented a vision transformer. To recognize static Indian sign language, proposed methodology archives noticeable performance over other state-of-the-art convolution architecture. The suggested methodology divides the sign into a series of positional embedding patches, which are then sent to a transformer block with four self-attention layers and a multilayer perceptron network. Experimental results show satisfactory identification of gestures under various augmentation methods. Moreover, the proposed approach only requires a very small number of training epochs to achieve 99.29 percent accuracy.

INDEX TERMS Transformer, deep learning, sign language, multi head attention, healthcare, mental health.

I. INTRODUCTION

A communication medium consists of hand gestures and the most structured and organized language to effectively communicate for impaired people. Sign language is a collection of various gesture-generation techniques. Sign language is a more effective method of communication than leap moment identification or writing a message. Sign language is vast and consists entirely of gestures to properly comprehend messages. Sign language is not just a gesture using fingers and palms; it involves visual cues through the eyes, face, mouth, eyebrows, etc. Additional components, like facial expressions, involve expressing the complex meaning. Normal verbal language is much more creative and cultivated than normal verbal language. The artistic spirit of life is given by the hand moment, body, and facial expression. Although the sign language can be simple and professional, it can also be an animated way to communicate, even though the sign language is very formal.

Sign language recognition is an area of research that involves pattern matching, deep learning, computer vision, natural language processing, and a design module or algorithm to identify sign language. It can be extended further to human-computer interaction without a voice interface. This system belongs to multidisciplinary content and the approach can be considered as a part of the Sign Language System.

There are around 300 sign languages used around the world [1]. The numbers don't have some level of confidence because day by day some countries immerge with their own sign language. American, British, and Chinese sign languages are the most widely used worldwide [1]. There is a very huge diversification in sign language, which varies from region

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

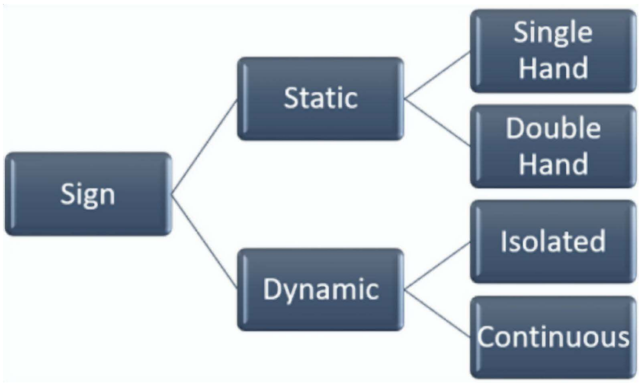


FIGURE 1. Formal classification of sign language.

to region. It involves several local or conventional subsets of the language. The diversity may lead to different flows of gesture singing, expression, jargon and the formation of gestures. As certain English words are spoken differently in different parts of one country, different accents and distinct dialects may be present in sign language. Indian sign language is the most widely used language in the South Asian regions [2].

Sign language is classified based on static and dynamic or involving manual and non-manual body parts. This classification can be helpful to researchers and designers of sign language recognition systems. We must combine both sign components to design a robust or real-time sign language recognition system. As shown in figure 1, sign language can be divided into two basic categories: static and dynamic. Static signs can be generated using one hand or two hands, while dynamic signs are further divided into two subcategories as isolated and continuous. To include emotional substances, dynamic signs can be further divided by the involvement of non-manual body parts. Non-manual body parts can be eyes, head movements, leaps, and eyebrows [3].

The main contributions of the work as, i) Proposed Transformer-based DeepVisionTransformer to recognize Indian sign language. ii) Evaluate the proposed model with a very small number of learning cycles (5 epochs), which is tiny compared to other state-of-the-art sign recognition models

The rest of the article organized as section II contains literature study relevant to sign language recognition. Section III includes a detailed description of Vision Transformer and Vision Transformer for image recognition. Section IV contains details of the proposed methodology. Section V contains details of the experiment and result. Section VI present discussion and conclusion of proposed work.

II. RELATED WORK

Rokade [4] proposed a methodology for automatically recognising fingerspelling in the Indian sign language. Input the sign image first to perform segmentation based on skin colour to detect the sign's shape. The detected area is converted into a binary image. Furthermore, a Euclidian distant transformation is applied to the binary image. After the feature extraction using Hu's moment, classification is done with ANN and SVM. The accuracy was 94.37% with SVM and 92.12% with ANN over 13 features [4]. The author has found good accuracy with ANN even with a smaller number of features set. The author has used a black background image of the letter (26class) with dimensions of 320*240.Video-based Indian signs are used to recognize them by the proposed system. Katoch et al. [5] present a technique that uses the "Bag of Visual Words" model (BOVW) to recognize Indian sign language letters and digits. The proposed methodology uses segmentation based on skin color and background subtraction. The authors used histogram-based sign mapping. At the end, CNN and SVM were used for classification. The author has also developed a GUI to make access easier. The author has used a custom dataset of more than 36,000 images to recognize Indian sign language. Over the dataset mask generates binary and canny edges to extract the feature with SURF. The proposed methodology with SVM and CNN found 99.17% and 99.64% accuracy, respectively [5]. Shenoy et al. [6] proposed a static hand pose for Indian sign language recognition. Video frames are captured from a smartphone and transmitted to a server for processing. The author has used skin color segmentation for hand detection and tracking. Feature extraction uses a grid base technique to represent hand gestures in a feature vector. The author has used KNN for static hand pose (alphabet and number) classification, while HMM was used to classify other gestures of Indian sign language with an accuracy of 99.7% and 97.23%, respectively [6]. The author used a custom dataset of over 24624 images for the experiment. De Coster et al. [7] proposed a sign language recognition methodology over the Flemish Sign Language corpus. The author has used OpenPose feature extraction and end-to-end learning with CNN, and applied a multi-head attention approach to isolated sign recognition. Over the class of 100 signs, 74.7% accuracy has been obtained as a state-of-the-art result over the Flemish Sign Language Corpus. The author introduces the Multimodal Transformer Network with Pose LSTM and Pose Transformer, especially self-attention for sign language recognition [7]. Mannan et al. [8] proposed Hypertuned DeepCNN for American Static sign language, author has used data augmentation to create more number of learning data sample, as deep learning model accuracy will increase with more samples for the training process. The proposed architecture follows conventional CNN with tuned hyper that parameters able to achieve 99.67% accuracy with 20 epochs. Zakariah et al. [9] proposed CNN based architecture for Arabic letter sign recognition. The authors generated 160000 images from 32000 images with data augmentation, which helps to consider different brightness and angular scenario. EfficientNetB4 method used for simulation. The authors also modified existing EfficientNet by adding one fully connected dense layer Author used a standard ArSL2018 dataset with 32 classes and get 95% of accuracy in 30 epochs. Kamruzzaman [10] proposed CNN based method for Sign language detection. Authors have proposed ResNet50 and MobileNetV2 based methodology for

TABLE 1. Comparative analysis of training phase prospective with state-of-the-art results in static sign recognition.

Author	Class	Methodology	Learning Parameter	Feature Extractor	Result
Rokede Y. et al [4]	26	SVM	13 Feature sets	HU's Moment	92.12%
Rokede Y. et al [4]	26	ANN	13 Feature sets	HU's Moment	94.37%
Kachot et al. [5]	36	CNN	50 epochs	SURF	99.64%
De Coster M [7]	100	Multimodal transformer	100 epochs	–	74.70%
Mannan, A [8]	25	DeepCNN	20 epochs	–	99.67%
Zakariah, M [9]	32	EfficientNet B4	30 epochs	–	95%
Alnuaim, A [10]	32	ResNet50 + MobileNetV2	10epochs	–	98.20%

Arabic sign language recognition. ResNet50 and MobileNetV2 simulated separately on 32 classes of ArSL2018 dataset. A combination of ResNet50 and MobileNetV2 can able achieve 98.2% accuracy with 10 epochs. Rathi et al. [11] proposed deep learning based sign language recognition model. Authors have used 2-level ResNet50 architecture to recognize sign language, authors have used 36 classes of the American Sign Language dataset form Massey university [12] having approx. 70 RGB images. Proposed 2 level ResNet50 methodology archive 99.03% accuracy. S. Jiang, et al. proposed skeleton aware multi-model for sign language recognition [13]. The authors used hand detectors with a pose estimator to extract hand key points. Methodology introduces the sign language GraphCN(SL-GCN). As a result, proposed methodology archive 98.425 of accuracy over RGB images. Roman Tongi, introduced a transfer learning based methodology for sign language recognition [14]. The methodology proposes how transfer learning can be applied to SLR using inflated 3D convolution neural network. American Sign Language (MS-SAL) and German Sign Language Dataset (SIGNUM) were used for simulation, and archive compatible result. In 2017, Google Brain researchers proposed an encoder and decoder network architecture based on attention mechanisms. The author used this transformer architecture to translate the language. For a model architecture that completely relies on an attention mechanism, foregoing recurrence, to identify global dependencies between input and output. Experiment with 8 GPUs (P100). We obtain state-of-the-art results for English to French translations [15].

By following outstanding performance on a language task, the transformer opens up a new dimension for computer vision problems. Use a transformer for image classification. Attention can be implemented in conjunction with a convolution network in computer vision. The proposed method takes an image as input and does not extract any features. Instead, convert the image into patches, and the sequence of patches serves as an input matrix to the transformer's encoder layer. Further classification to be done with the MLP. The authors have introduced three variants of ViT, as Base, Large, and Huge, having 12, 24, and 32 layers, respectively [16]. Table 1 shows the comparative analytics of static sign language detection with state-of-the-art deep learning models, with parameter details and the number of classes.

III. MATERIALS AND METHODS

The proposed architecture used a transformer-based method for static Indian sign recognition, multihead self-attention is proposed in the encoder phase of the transformer, the detail study of vision transformer and multi-head attention as follow.

A. VISION TRANSFORMER

The emergence of Vision Transformer (ViT) strongly competes with the CNN, the state-of-the-art of computer vision, so commonly utilized in several image recognition tasks. The ViT models surpass the convolutional neural networks (CNNs) in terms of computational capabilities, efficiency, and accuracy [17]. In the field of natural language processing, transformer architectures have state-of-the-art performance standards. Only a few use cases are included in the field of computer vision because attention is used in association with convolutional networks (CNNs) or can be used as a substitute for certain convolution features while preserving their original composition. The transformer encoder detaches these dependencies of the CNN, and the standard transformer architecture can be directly applied to the sequences of image patches, and it works surprisingly well and accurately over image classification tasks.

Initially, the transformer was introduced for language processing tasks. The trans-former used attention mechanisms rather than convolution layers. The design of a trans-former consists of an encoder and a decoder. Both of them involve self-attention and feed-forward mechanism. The transformer can be applied to various computer vision tasks by performance capabilities. The transformer performs better than other convolution methodologies in various computer vision task [18]. The special use of a transformer in a computer vision task is known as a vision transformer (ViT). ViTs (different variants of ViT) achieve promising and remarkable results in computer vision tasks. ViT has two major benefits: 1) self-attention mechanism, where the model reads a long range of input seeds (tokens) in a universal context. 2) The ability to train on large tasks.

Figure 2 depicts the ViT design, in which first images are converted into patches in accordance with the model design. Then patches directly feed into the linear projection layer. In the second stage, the patch embedding process is performed. The class token has been added to the sequence of the embedded patches. Thus, the size of patches increases by one. Embedded patches are also added with positional embedding to the memory positional sequence of patches. Finally, patch embedding and positional encoding with a

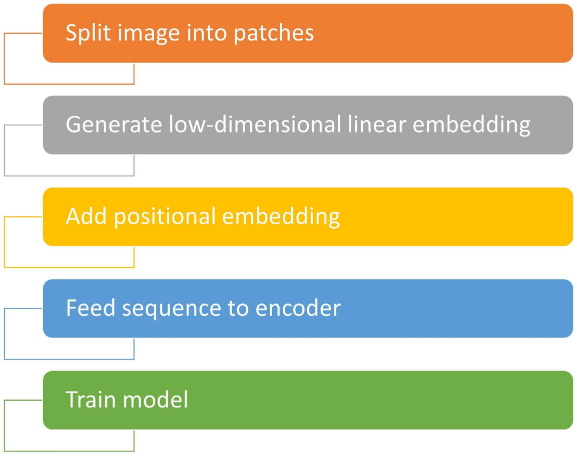


FIGURE 2. Visual transformer process sequence for image classification in an encoder.

class token are fed to the encoder layer as the first transformer layer. Encoding is the most important component of a transformer, especially in ViT. It contains two major components, Multi Head Self-Attention (MHSA) and Multi-Layer Perceptron (MLP). The embedded input is normalized through the normalization layer. A normalized value is used to obtain Query (q), Key (k), and value (v) as a matrix, as shown in equation (1). The MHSA module executes the following equation to achieve attention operation inside the encoding [19]. Finally, the attention layer's output is fed to the feed forward layer, which generates the encoder's final output.

$$Attention(q, k, v) = softmax(q * kk^T/\sqrt{d_k}) * v. \qquad (1)$$

B. VIT FOR IMAGE CLASSIFICATION

The vision, to vision transformer for image recognition task was introduced in "An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale" by Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov et al. and successfully trained on ImageNet, attaining good results compared to the convolution network [15]. The transformer encoder receives an input image, generates fixed size non-overlapped patches from these images, and then linearly embeds the sequence of patches. The class/token is embedded to represent an entire image, which can be needed at the classification phase. The author also adds the resulting vector's absolute position embedding and set sequence to the pure transformer encoder. The original image resolution and patch resolution used in training and fine tuning are reflected at every transformer layer checkpoint; however, any pre-train model can be used for the same. The author does this to improve the accuracy and predictive power of the transformer, because each head has its own way of internal representation and computation of input, each head can manage to understand the relationship between patches in sequence (i.e., collective, shared knowledge). Any relationship information among the patches missed by one head is highly likely to be covered by another head [20].

IV. PROPOSED ARCHITECTURE DESIGN

The proposed multihead self-attention based architecture contains three major parts: patch embedding, feature extraction and a classification head. Stacked encoders are a core part of the simulation because of the feature extraction. Algorithm 1 represent pseudo code for proposed methodology.

Algorithm 1 Sign Recognition

```
Input: RBG sign image
Output: Sign value of image
procedure data_input(img)
    x_train, x_test, y_test, y_train ← split(img, lable, split_ratio)
    return x_train, x_test, y_test, y_train
end procedure
procedure DATA_AUGMENTATION(img)
    img ← IMG.NORMALIZATION()
    img ← IMG.RESIZE(72, 72)
    img ← IMG.RENDOMFLIP(Horizontal)
    img ← IMG.RENDOMROTATION(0.02)
    return img
end procedure
procedure PATCHES(img,patch_size)
    patches ← IMG.RESHAPE(img, extract_patch, batch_size, num_patch)
    return patches
end procedure
procedure PATCHENCODER(patches)
    patches ← IMG.RESHAPE(img, extract_patch, batch_size, num_patch)
    position ← IMG.RANGE(0, num_patch, delta)
    encoded ← SELF.PROJECTION(patch) + SELF.POSITION_EMBEDDING(position)
    return encoded
end procedure
for each_img do
    inputs ← DATA_INPUT(img)
    augmented ← DATA_AUGMENTATION(inputs)
    patches ← PATCHES(augmented, patch_size)
    encoded_patches ← PATCHENCODER(patches)
    features ← TRANSFORMER(encoded_patches)
end for
classification ← MLP(features)
return class_sign
```

A. PATCH EMBEDDING

Initially 2D images convert into 1D sequences of embedding tokens. Images (Xi) are reshaped as, $X \in R^{(H*W*C)}$ where c is the number of channels as 3, and (H, W) consider a weight and height as (72 × 72). These images are converted into the sequence of flattened 1D patches with the shape of $(N, P^2 * C)$, where N is the total number of patches and (P, P) is the dimension of the patch. A positional embedding tensor (Epos) with shape of (N, D), learns 1D positional information of each patch and generates the spatial representation of the patches.

B. MULTIHEAD SELF-ATTENTION

The multihead mechanism learns embedding vectors with different aspect. The multihead attention (MHA) is included in each of the 8 layered stacked transformers. The hidden state divided into n=4 heads to generate n feature tensor. Each self-attention head has three trainable matrices (q, k, v), represented in equation 2. Every (four) heads have an attention tensor that can calculate as equation 1. softmax operation

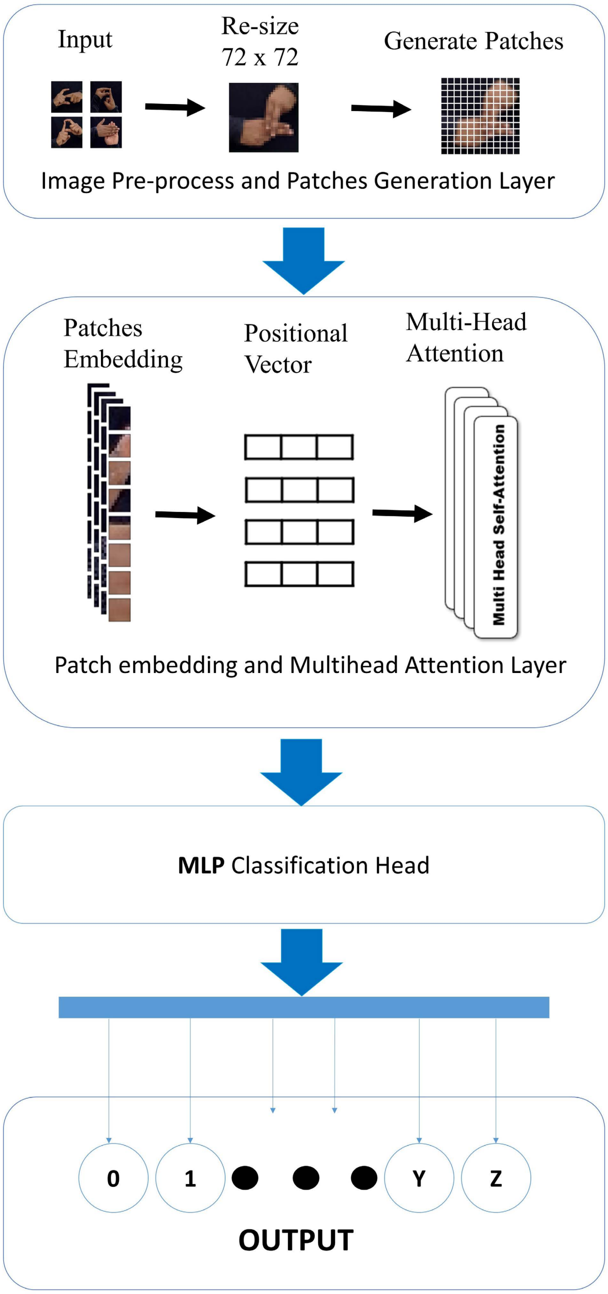


FIGURE 3. Proposed Transformer encoder architecture for static Indian sign language recognition.

gives the attention score for every attention head. Farther self-attention matrices can be calculated as dot products of A and v (equation 3) [16], and concatenated features of the tensor can be generated with equation 4 [16].

$$[q, k, v] = zU_{qkv} \quad (2)$$

$$SA(z) = Av \quad (3)$$

$$MSA(z) = [SA_1(z); SA_2(z); SA_3(z); \ldots..; SA_n(z)]U_msa \quad (4)$$

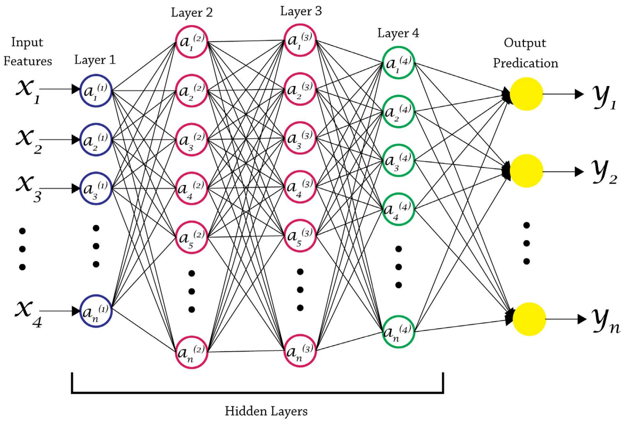


FIGURE 4. MLP classifier with four hidden layers.

C. MLP CLASSIFICATION

An output tensor of multihead is added to the residual connection, which is projected by point wise feed-forward network with two linear layers with ReLU activation in between. Each layer uses different weights (W1, W2) and bias (b1, b2) as shown in equation 5.

$$FFNN(x) = ReLU(W_{1x} + b_1)W_2 + b_2 \quad (5)$$

Proposed methodology uses MLP [21] classifier with four hidden layers of vary in size to perform classification over Multihead transformer network as shown in figure 4. classification network is defined as $x_n * a1_n * a2_n * a3_n * a4_n * y_n$ were x_n represent input feature vector, $a(x)_n$ represent neurons in respective hidden layers, y_n represent output class prediction. Proposed methodology uses MLP classifier based on capabilities like i) ability to learn in complex and non-linear networks ii) Generalization ability can be improved iii) MLP learns independently from input variable size [22]. Farther Adam Deep learning optimizer has been used, which inherit the feature of RMSprop and AdaGard [23]. Parameter of classifier were set to improve model performance.

The authors have proposed a transformer-based encoder model to recognize static Indian sign language from an image-based dataset. Initially, the dataset was divided into train validation splits with a 0.2 splitting factor, and the read resized image was 72 × 72. Resize image converted to the same size as the non-overlapped patches. The proposed methodology creates 144 patch form input images, as shown in figure 3. The sequential embedding layer creates a patch sequence, which is further combined with the positional vector. The output of the positional embedding layer is fed to multi-head attention. The proposed architecture uses six self-attention layers, and an appropriate tensor is managed at the positional embedding layer. The classification has been performed in the MLP head as a single layer of fine-tuned time. The proposed methodology can archive validation accuracy of 99.29% with only five epochs. A small amount of training

TABLE 2. Specification of dataset used in simulation.

Dataset Name	No of Class	Average image per class	Original Resolution
Indian Sign Language [24]	36	1000	128 x 128
Indian Sign Language [25]	36	1200	250 x 250
American Sign Language [30]	36	841	400 x 400
Bangla Sign Language [31]	33	654	171 x 166

can result in a high recognition rate. As per table 1, the proposed methodology can recognize static signs with little training and high accuracy.

V. EXPERIMENT SETUP

A. DATASET

The dataset used in the simulation is prepared from collection of publicly available Indian sign language dataset (static) [24], [25], which includes gesture of numbers (0–9) and alphabet. Dataset consist of RGB images of total 36 classes with more than 1000 images per class to improve data generalization augmentation has been done. Table 2 shows the characteristics of dataset used in this experiment.

B. DATA AUGMENTATION

Data augmentation is mainly used for sample balancing and improving training sample variability. Data augmentation is also significant for transformer based framework because the huge amount of data is essential for model training. Using different augmentation techniques like flipping, cropping, rotation, etc., authors simulate horizontal flipping, colour space transformation, random zoom with 0.2 weight and height and slight rotation from -1 to -10 to train the model on more generalized data. Figure 5 shows an overview of data set after augmentation.

C. IMPLEMENTATION DETAILS

The authors have worked on a standard Static Indian Sign Language dataset [24], [25]. During this study, the authors implemented a modified transformer using TensorFlow-Keras. The proposed methodology has achieved 99.29% accuracy over the 36 image-based Indian sign language classes. The proposed methodology uses a static Indian sign language dataset [22] of images with a size of 480×320 pixels with 3 (RGB) channels. The dataset is split into 2 parts (train-test) with a 0.2 splinting rate as for training and testing. Initially 72×72 images are converted in patch size of 6×6, total 144 patches as $((image\_size/patch\_size)^2)$ will be created for every image. The position embedding tensor of the patches will be used as the encoder input. Farther tensor will pass through two normalized layers with activation (ReLU) and MLP classification head. The performance of the proposed methodology was evaluated by three different experiments. The parameters used in training are fine-tuned, like the optimizer number of layers and activation. Precision [26], recall [27] and f1-score [28] were calculated as equations 6, 7 and 8, respectively, where TP and FP are the number of true and false positives, respectively

TABLE 3. Specification of dataset used in simulation.

Dataset	RGB	Background	Accuracy	Precision	Recall	F1-Score
ISL [25]	YES	Dynamic	95.67	0.96	0.96	0.96
ISL [24]	YES	Static	99.29	0.99	0.99	0.98

$$Precision = TP/(TP + FP) \tag{6}$$

$$Recall = TP/(TP + FN) \tag{7}$$

$$F1 - Score = 2 * (Precision * Recall/(Precision+Recall)) \tag{8}$$

The proposed methodology has achieved significant accuracy using a smaller number of attention layers in the encoding component of the transformer as well as a very small number of learning cycles. All the results were taken on a personal computer with an Intel Core i7 and 16 GB of RAM only. Jupyter Lab is used to implement the proposed methodology. Figure 8 represents a heat-map for the class wise recognition of static sign language.

D. RESULT OF PROPOSED METHOD

In this simulation, we evaluated different combinations of state-of-the-art classification networks with different classifiers. Furthermore, the Author has also simulated proposed methodology with different classifiers and train test split ratio.

1) EXPERIMENT WITH DIFFERENT BACKGROUND

Table 3 shows result comparison of The Indian sign language dataset with and without static background, in both cases training-testing parameters will be constant as five epochs and 80-20 train-test splitting ratio has been taken for simulation. Data augmentation was performed in both the scenario. The outcome of this experiment to the analysis of the proposed methodology can recognize sign gestures over various backgrounds (vary from sign to sign). The different appearances of the background do not depend on the environmental factor and resolution of camera.

2) EXPERIMENT WITH DIFFERENT TRAIN–TEST SPLIT RATIO

In this experiment, the we simulated the proposed methodology's performance over different train-test split ratios as 80-20, 70-30, 60-40 shown in table 4. The reduction in the training dataset slightly effect on the recognition rate.

3) EXPERIMENT WITH DIFFERENT SELF-ATTENTION HEAD

Author also experiment with different numbers of self-attention head, as figure 6 shows that classification accuracy differs with the number of attention head. This study aims

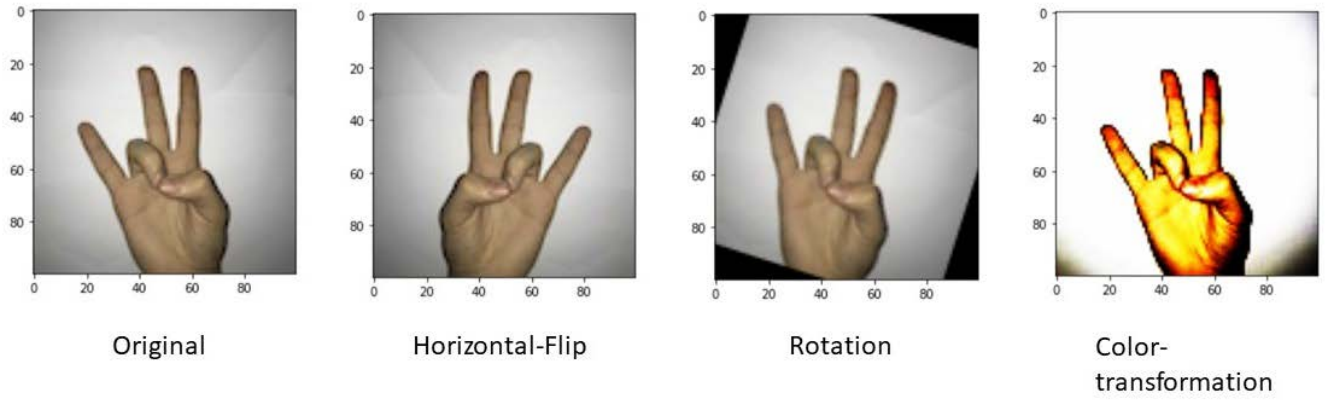


FIGURE 5. Sample of augmented data with horizontal flip, rotation and colour space transformation.

TABLE 4. Specification of dataset used in simulation.

Train-test Split ratio	Accuracy	Precision	Recall	F1-Score
80-20	99.29	0.99	0.99	0.98
70-30	99.04	0.99	0.99	0.98
60-40	98.71	0.98	0.98	0.98

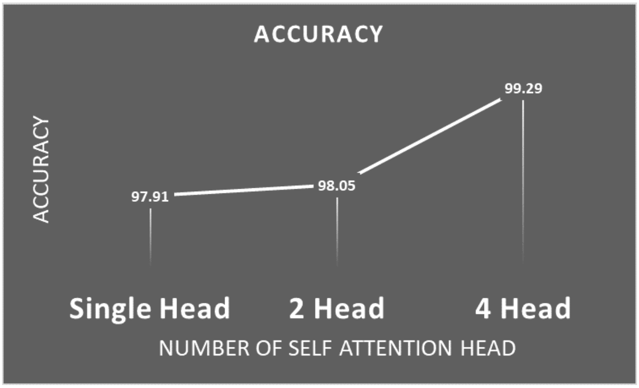


FIGURE 6. Accuracy comparison with multiple self-attention head.

to know how classification accuracy depend on the number of self-attention head. It also observes that from one head to 2 head, there is no major change in accuracy, whereas 2 to 4 head shows exponentially changes in accuracy.

4) EXPERIMENT WITH DIFFERENT CLASSIFIER

Figure 7 present the comparative analytics of different state-of-the-art classifiers over the Indian sign language dataset [24]. VGG16, VGG19, Inception V3 and ResNet-50 were taken for convolution with different classifiers.

E. PERFORMANCE MEASURES

Several standard metrics for performance evaluation like accuracy, classification error and precision have been considered for model computation performance measurement. Accuracy can be considered an indicator of the model's performance across all classes. The precision can be calculated as the ratio of the total number of positive samples identify correctly over the total number of samples classified as positive. Classification errors can be defined as missing of classification accuracy and error in the classification instant. Figure 8 shows the classification results of all 36 classes as a part of class wise performance analytics. Heat map represent erroneously classify signs. These three performance metrics are used to better understand the model performance with the existing model to identify the significant performance of Transformer for sign language recognition. Figure 9 shows the result comparative analysis of Indian Sign Language Dataset with augmentation and MLP classifier with ReLU activation over other state-of-the-art deep learning methodology of gesture recognition. Augmented ISL datasets have been tested over CNN(core), Fast-R-CNN and Adaptive CNN. Five training cycles (epochs) have been taken as static parameters for comparisons, figure 9 also represent graphical representation of the classification report over tested deep learning models.

TABLE 5. Performance analysis of the proposed architecture on other static sign language dataset.

Dataset	Accuracy	Precision	F1-Score	Classification Error
Indian Sign Language [25]	99.29%	0.99	0.99	0.71
American Sign Language [29]	98.31%	0.98	0.99	1.69
American Sign Language (Augmented) [30]	97.68%	0.98	0.98	2.32
Bangla Sign Language [31]	98.79%	0.99	0.99	1.21

F. DISCUSSION

In this article sign language recognition is considered for static Indian sign language. The proposed methodology may assist impaired people in communicating with other

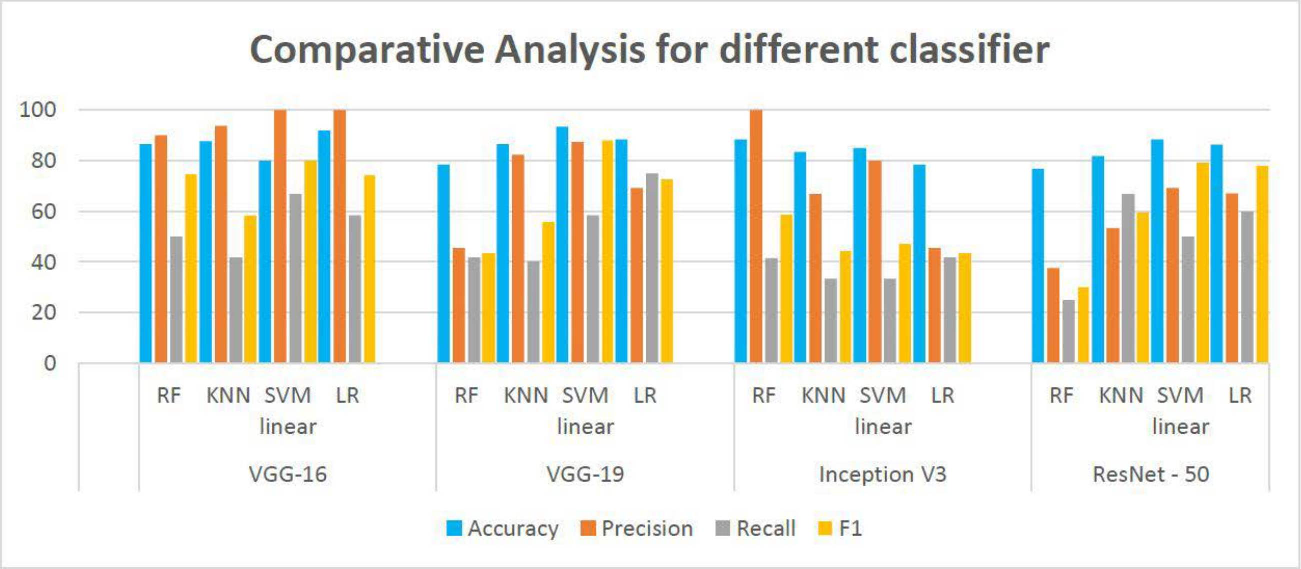


FIGURE 7. Different classifier comparative analysis of state-of the-art image based classification network. (RF=Random Forest, KNN = K-Nearest Neighbour, SVM= Support Vector Machine, LR = Logistic Regression.)

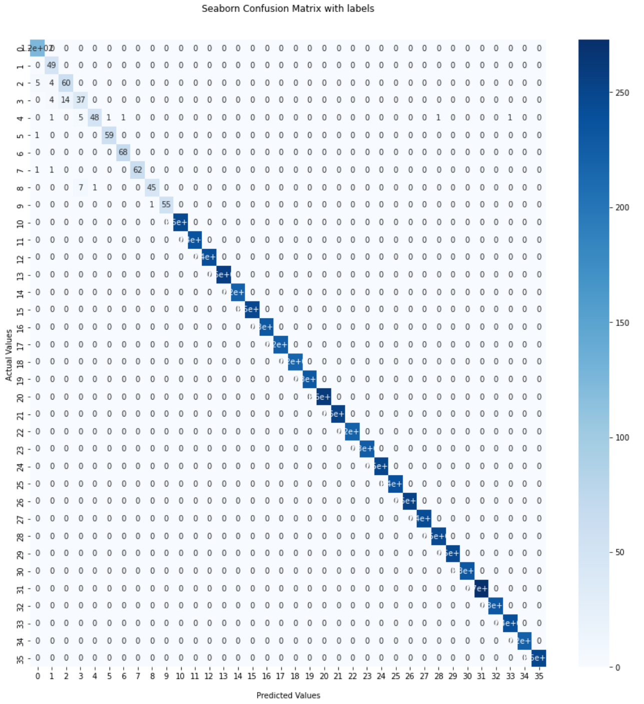


FIGURE 8. Heat map graph for static Indian sign recognition.

normal people. The dataset used in the study is a static ISL dataset having signs of digit and alphabet in the context of the Indian community. The proposed methodology is able to achieve very good accuracy as 99.29% with a higher number class as 36 and very small number of training cycles as five epochs. There is no need for data pre-processing while working with the transformer encoder. Multihead attention

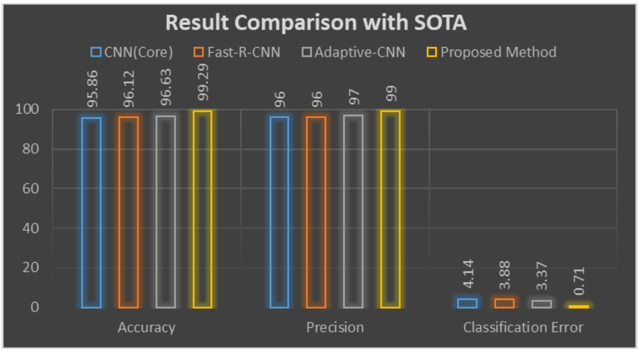


FIGURE 9. Performance analysis of existing model on static Indian sign language (ISL) dataset.

helps the model to improve the performance compared to other standard transformer model. Table 5 shows the performance of the proposed methodology over other standard static sign language datasets, such as American Sign Language (ASL) with and without augmentation and Bangle Sign Language. The proposed methodology also performs effectively on other standard datasets with the same number of training cycles as five epochs.

VI. CONCLUSION

Sign language recognition systems have lots of potential applications in the field of human-computer interaction. A vision-based static sign gesture recognition system is essential to reduce the communication gap between normal people and visually impaired people. The proposed methodology presents transformer-based sign language recognition for static signs. A multi-head attention-based encoding framework can achieve good accuracy with a very small number of training layers and epochs. A framework with a tiny training process can also find good accuracy over a large set of classes. The multi-head encoding framework in the Transformer broke up the recognition rate of gesture and sign language recognition as a part of human-computer interaction applications. The Proposed methodology can also detect images with augmentation like different angular position and different brightness levels. Multihead based transformers are successful for static sign recognition, for more advancement proposed model can be modified for isolated and continuous sign language detection. Father transformer based methodology can proceed to identified sign gesture form isolated video of sign language, and farther extended to recognize continues sign language.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of Prince Sultan University for paying the APC of this publication.

REFERENCES

[1] W. Bright, "Ethnologue: Languages of the world ed. by Barbara F. Grimes, and: Index to the tenth edition of ethnologue: Languages of the world ed. by Barbara F. Grimes," *Language*, vol. 62, no. 3, p. 698, 1986.

[2] I. Meir, "Review of Zeshan (2000): Sign language in Indo-Pakistan: A description of a signed language," *Sign Lang. Linguistics*, vol. 3, no. 2, pp. 263–267, Dec. 2000, doi: 10.1075/sll.3.2.10mei.

[3] 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.

[4] Y. Rokade, "Indian sign language recognition system," *Int. J. Eng. Technol.*, vol. 9, no. 3, pp. 189–196, 2017.

[5] S. Katoch, V. Singh, and U. S. Tiwary, "Indian sign language recognition system using SURF with SVM and CNN," *Array*, vol. 14, Jul. 2022, Art. no. 100141.

[6] K. Shenoy, T. Dastane, V. Rao, and D. Vyavaharkar, "Real-time Indian sign language (ISL) recognition," in *Proc. 9th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2018, pp. 1–9.

[7] 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. 3441–3450.

[8] A. Mannan, A. Abbasi, A. R. Javed, A. Ahsan, T. R. Gadekallu, and Q. Xin, "Hypertuned deep convolutional neural network for sign language recognition," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–10, Apr. 2022.

[9] M. Zakariah, Y. A. Alotaibi, D. Koundal, Y. Guo, and M. M. Elahi, "Sign language recognition for Arabic alphabets using transfer learning technique," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–15, Apr. 2022.

[10] M. M. Kamruzzaman, "Arabic sign language recognition and generating Arabic speech using convolutional neural network," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–9, May 2020.

[11] P. Rathi, R. K. Gupta, S. Agarwal, and A. Shukla, "Sign language recognition using ResNet50 deep neural network architecture," in *Proc. 5th Int. Conf. Next Gener. Comput. Technol. (NGCT)*, 2020.

[12] A. L. C. Barczak, N. H. Reyes, M. Abastillas, A. Piccio, and T. Susnjak, "A new 2D static hand gesture colour image dataset for ASL gestures," *Handle Proxy*, vol. 15, pp. 12–20, Jan. 1970. Accessed: Oct. 18, 2022. [Online]. Available: http://hdl.handle.net/10179/4514

[13] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multimodal sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3413–3423.

[14] M. Varsha and C. S. Nair, "Indian sign language gesture recognition using deep convolutional neural network," in *Proc. 8th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jul. 2021, pp. 193–197.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin, and L. Kaiser, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[17] G. Boesch. (Aug. 23, 2022). *Vision Transformers (ViT) in Image Recognition—2022 Guide*. Accessed: Oct. 18, 2022. [Online]. Available: https://viso.ai/deep-learning/vision-transformer

[18] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[19] K. Islam, "Recent advances in vision transformer: A survey and outlook of recent work," Aug. 2022, *arXiv:2203.01536*. Accessed: Oct. 18, 2022.

[20] R. Anand. (Jul. 2, 2021). *Vision Transformers Explained*. Accessed: Oct. 18, 2022. [Online]. Available: https://blog.paperspace.com/vision-transformers

[21] B. Krose and P. V. D. Smagt, *An Introduction to Neural Networks*. Princeton, NJ, USA: Citeseer, 1993.

[22] H. K. Gajera, D. R. Nayak, and M. A. Zaveri, "A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104186.

[23] J. Singh and R. Banerjee, "A study on single and multi-layer perceptron neural network," in *Proc. 3rd Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Mar. 2019, pp. 35–40.

[24] P. Arikeri. (Jun. 4, 2021). Indian sign language (ISL). Kaggle. Accessed: Oct. 18, 2022. [Online]. Available: https://www.kaggle.com/datasets/prathumarikeri/indian-sign-language-isl

[25] D. R. Kothadiya. Deepkothadiya/STATIC_ISL: Static Indian sign language dataset having sign of digit and alphabet. GitHub. Accessed: Oct. 19, 2022. [Online]. Available: https://github.com/DeepKothadiya/Static_ISL

[26] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun. 2015, pp. 1–6.

[27] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using real-sense," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Jul. 2015, pp. 166–170.

[28] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 572–578.

[29] GrassKnot. Grassknoted/unvoiced: American sign language to speech application. GitHub. Accessed: Oct. 18, 2022. [Online]. Available: https://github.com/grassknoted/Unvoiced

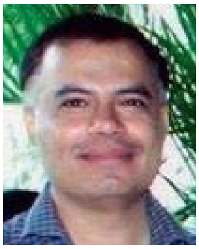
[30] A. Thakur. (May 1, 2019). American sign language dataset. Kaggle. Accessed: Oct. 19, 2022. [Online]. Available: https://www.kaggle.com/datasets/ayuraj/american-sign-language-dataset

[31] S. M. Rayeed. (Aug. 8, 2021). Bangla sign language dataset. Kaggle. Accessed: Oct. 18, 2022. [Online]. Available: https://www.kaggle.com/datasets/rayeed045/bangla-sign-language-dataset

DEEP R. KOTHADIYA received the bachelor's and master's degrees in computer science and engineering from Gujarat Technological University. He is currently pursuing the Ph.D. degree with the Charotar University of Science and Technology (CHARUSAT). He is also working as 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 CHARUSAT. He has already published three research papers, including one SCI indexed paper. He is also a Technical Reviewer of *International Journal of Computing and Digital Systems* (Scopus).

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

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, University Teknologi Malaysia (UTM), Malaysia, in 2012. She is currently working as an Associate Professor with the College of Computer and Information Sciences, Prince Sultan University (PSU), Riyadh, Saudi Arabia. She has published more than 100 publications in high ranked journals. Her research interests include bioinformatics, data mining, and classification. She won the Best Student Award at the Faculty of Computing, UTM, in 2012. She was awarded the Best Research of the Year Award at PSU, from 2013 to 2016. Due to her excellent research achievement, she is included in Marquis Who's Who (S & T) 2012. She is also an editor of several reputed journals and on panel of TPC of international conferences.



AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. degree (Hons.) from the Faculty of Computing, Universiti Teknologi Malaysia, in 2010, with a specialization in forensic documents analysis and security. He was a Postdoctoral Researcher with the Faculty of Computing, Universiti Teknologi Malaysia, in 2011. He is currently a Senior Researcher with the Artificial Intelligence and Data Analytics Laboratory, CCIS, Prince Sultan University, Riyadh, Saudi Arabia. He is also a PI in several funded projects and also completed projects funded from MOHE Malaysia, Saudi Arabia. He is the author of more than 200 ISI journal articles and conferences. His research interests include data mining, health informatics, and pattern recognition. He received the Rector Award for the 2010 Best Student in the university.

SAEED ALI BAHAJ received the Ph.D. degree from Pune University, India, in 2006. He is currently an Associate Professor with the Department of Computer Engineering, Hadramout University, and also an Associate Professor with Prince Sattam bin Abdulaziz University. His research interests include artificial intelligence, information management, forecasting, information engineering, big data, and information security.