IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 3, June, 2022, pp. 823-833
Journal homepage: http://iieta.org/journals/ts

A Novel Hyperparameter Optimization Aided Hand Gesture Recognition Framework Based on Deep Learning Algorithms

Abdullah Asim Yilmaz

Computer Engineering Department, Ankara University, Golbasi 06830, Ankara, Turkey

Corresponding Author Email: aayilmaz@ankara.edu.tr

https://doi.org/10.18280/ts.390307

Received: 17 March 2022
Accepted: 13 May 2022

Keywords:
hand gesture recognition, deep neural network, transfer learning, deep learning, genetic algorithm

ABSTRACT

The recognition of hand gestures in cluttered or complex environments is a vital research area in the human-computer interaction and computer vision fields due to its various potential applications, such as hand action analysis, driver hand behaviour monitoring, virtual reality, pose estimation, human action recognition, and sign language recognition. In order to create more reliable and efficient algorithms in this research field, various approaches have been suggested in recent years. However, a robust system is still elusive. For this reason, a new deep learning-based architecture for classifying hand gestures is suggested in this study; it is based on a hybrid model. The study makes two main contributions to the literature. The first is the creation of a new database for hand gesture recognition. The second is a novel hybrid architecture that combines two widely used pre-trained network models in an optimised manner, using a genetic algorithm for hyperparameter optimization. The proposed method comprises five main phases, namely, data acquisition, pre-processing, the design of the deep neural network architecture, hyperparameter optimization, the training of the proposed deep neural network architecture. The proposed method was tested on three comprehensive datasets. The experimental results reveal that the suggested method can effectively classify hand gestures with a high accuracy rate and that it outperforms the state-of-the-art methods in the literature.

1. INTRODUCTION

Sign language is a linguistic communication framework that consists of hand gestures that represent different words. Sign language differs from one region to another. According to the World Health Organization, more than 5% of the world's population, approximately 430 million people, suffer from hearing loss and deafness [1]. Sign language arose to allow deaf and mute people to communicate. Each of the different hand gestures is made by using the fingers and palms to communicate different meanings in sign language. Although deaf and mute people can communicate with sign language, most people with normal hearing do not know sign language and cannot understand it. In addition, if deaf and mute people do not know the same sign language, it is not possible for them to communicate. Interpreters can solve this problem; however, they are not available for all situations. At this point, automatic hand gesture recognition is an advantageous and practical competitive solution for breaking the communication barrier between deaf and hearing people; it could make communication with deaf people easier than ever. In daily life, people use gestures not only to embellish their communication but also to be better understood. Gestures can mean basic commands like stop, up, and down, or they can be used to point to objects; they can also be used for more complex communication or to communicate emotional signals between human beings [2]. According to Hall [3], gestures constitute 65% of the communication between people, and verbal communication constitutes only 35%. Therefore, the expression and communication potential of gestures cannot be underestimated. As Karam pointed out, hand gestures are used for communication more than other gesture types [4].

With the development of technology, we are surrounded by intelligent devices, such as robot cleaners, communication devices, automatic car interfaces, etc. When we use computer-based technologies, we need to interact with them. Thus, human-computer interaction is currently important. There are solutions that allow human beings to interact with computers; however, most of them require additional external devices, such as a keyboard, mouse, etc. At this point, hand gestures are a good alternative way to interact with a computer interface. For instance, such sensitive touchless interfaces in cars allow interior feature control using hand gestures without distracting the driver [5]. Besides the field of human-computer interaction, hand gestures are used in a variety of other areas, such as the control of robots [6], virtual reality technologies [7], the control of home automation systems [8], and virtual games [9]. In addition, hand gestures are mainly classified into two categories: dynamic and static. A static hand gesture is an immobile hand shape that remains immobile for some amount of time. On the other hand, a dynamic hand gesture comprises the dynamic motion of the fingers or hands, which can impart a specific meaning.

In recent years, deep learning methods have been widely used for different problems, such as natural language processing [10], object detection [11-13], speech recognition [14], human action recognition [15], malware classification [16], driving safety [17], medical applications [18, 19], facial emotion recognition [20], and graph-based applications [21, 22].The success of deep learning methods has attracted the attention of researchers who work on sign language recognition, and studies in this area have been carried out [23-25] to eliminate the shortcomings of existing hand gesture recognition approaches and increase model performance. However, a robust system is still elusive. The creation of robust systems is crucial in terms of reducing computational complexity and feature spaces, increasing the system accuracy rate, and overcoming hardware and time resource constraints for hand gesture recognition systems. Therefore, this paper suggests a new hybrid deep learning-based hand gesture recognition architecture. In the suggested method, hand gesture examples are collected from the LaRED and TinyHands datasets and a personal dataset. Hand gesture samples were first pre-processed. In this step, the hand gesture samples in video format were converted into grayscale sequential single images. After image acquisition was completed, hyperparameter adjustment was performed using the development set derived from the training set and our deep learning-based architecture. A genetic algorithm (GA), which is a heuristic search algorithm, was used for parameter tuning. After this, the method extracted high-level hand gesture features from hand gesture data by utilising the convolution layers of the suggested hybrid architecture. Finally, the system was trained by utilising the optimised hyperparameter set obtained using GAs. Overall, two popular deep learning models are merged in order to generate a hybrid model, which relies upon the transfer learning method. The obtained results showed that the suggested method can effectively extract distinctive features for each hand gesture type and that it can classify distinct hand gesture types with a high accuracy rate. The six major contributions of this paper are summarised as follows: First, a new personal hand gesture dataset is created. Second, a novel hybrid deep learning-based method is proposed for hand gesture recognition. Third, the proposed method uses GA methods to optimise its hyperparameters. Fourth, the proposed method utilises a new hybrid layer that involves two pre-trained models instead of one model. Fifth, the method is trained and tested on a personal dataset and two well-known hand gesture datasets. Finally, the method has a higher accuracy rate than other known methods, and it reduces feature spaces significantly.

The remainder of this paper is structured as follows: Section 2 gives a brief review of some related work concerning hand gesture recognition. Section 3 describes the proposed framework in detail. In section 4, the experimental results and discussion are presented and the hand gesture datasets are explained. Finally, we conclude the paper with a brief summary and discuss future work in section 5.

2. RELATED WORK

The main goal of hand gesture recognition is to solve the problem of recognising static and dynamic hand gestures in which the bare hand assumes various poses to reveal specific meanings. Hand gesture recognition plays an important role in many applications, such as sign language recognition for deaf and speech-impaired people [26, 27], driver hand monitoring and hand gesture commands for the reduction of driver distraction [28], in-air writing interaction [29], hand-object interaction in augmented and virtual reality environments [30, 31], and many others. In this section, we present a review of the literature concerning hand gesture recognition methods using conventional handcrafted features, as well as those using deep learning networks.

Rao et al. [32] proposed an improved hand gesture recognition system utilising a hidden Markov model. In this study, a Markov model is constructed for foreground fingers in hand gesture images. This model was employed in the training and testing modes of a binary classification approach. An accuracy rate of 90.6% was obtained by this model. Gupta et al. [33] used a Gabor filter along with a combination of principal component analysis (PCA) and linear discriminant analysis (LDA), which is followed by a support vector machine (SVM) for classification, in order to realise hand gesture recognition. In the proposed method, 15 Gabor filters was employed to reduce the complexity with better accuracy. Here, PCA was used to overcome the small sample size problem, while LDA was utilised for feature extraction and feature reduction in this study. Rahman and Afrin [34] utilised an SVM classification approach to classify the hand gesture images. Here, feature vectors were selected using the biorthogonal wavelet transform. Then, hand gestures were classified into ten categories by a multiclass SVM. The authors achieved an accuracy of 85.7%. Marium et al. [35] suggested a hand gesture recognition system employing a convexity algorithm approach. The authors implemented this filtering approach on hand gesture images with the same background and achieved an accuracy of 87.5%.

Despite their promising performance in hand gesture recognition, the above-mentioned conventional methods have various limitations, such as high model complexity and the need for an excessive amount of training data. Therefore, deep learning methods have been used in hand gesture recognition recently to overcome these limitations. Oyedotun and Khashman [36] suggested the use of stacked denoising autoencoders and convolutional neural networks to recognise 24 American Sign Language hand gestures obtained from Thomas Moeslund's gesture recognition database [37]. Devineau et al. [38] introduced a 3D hand gesture recognition approach based upon a deep learning model. They developed a new convolutional neural network model. Only hand-skeletal data was employed here, and no depth image was employed. The authors achieved accuracies of 91.28% and 84.35%, respectively, for cases with 14 and 28 gesture classes. Chevtchenko et al. [39] developed a new method which combines a convolutional neural network (CNN) and traditional handcrafted features. This method is evaluated on three comprehensive datasets. The authors obtained better results by using combinations of image representation and validation techniques. Wu et al. [40] suggested a semi-supervised hierarchical dynamic framework based on a hidden Markov model for simultaneous gesture segmentation and recognition. In this approach, high-level spatiotemporal representations were learned by using deep neural networks suitable for the input modality, contrary to conventional methods. In addition, a 3D convolutional neural network was utilised to manage and fuse batches of depth and RGB images, while a Gaussian-Bernoulli deep belief network was used to handle skeletal dynamics. The authors obtained a Jaccard index score of 81.0% on the ChaLearn LAP gesture spotting challenge.

Despite the success of the aforementioned conventional and deep learning-based methods, there is still a lack of highly accurate approaches for recognising hand gestures in cluttered or complex environments. In this paper, we propose a novel hybrid deep learning architecture for recognising hand gestures effectively. In the proposed method, hand gesture samples were first collected from comprehensive datasets. Second, hand gesture samples were pre-processed. Third, hyperparameter adjustment was carried out using the development set derived from the training set and our deep learning-based architecture. In this step, a GA was used for parameter tuning because it had the best accuracy rate. Fourth, the features were extracted using pre-trained networks. Lastly, the training phase was realised by utilising the optimised hyperparameter set obtained using GAs and hand gesture training datasets and then classification is performed with utilizing softmax classifier.

3. PROPOSED METHOD

The proposed hand gesture recognition framework is presented in this section. Our proposed hand gesture recognition framework based upon deep learning methodologies includes a hybrid deep neural network architecture with hyperparameter optimization support for hand gesture recognition. The methodology of the proposed system, illustrated in Figure 2, is comprised of five main steps. First, the collection of the hand gesture data is accomplished by using a personal dataset and two exhaustive datasets in the input phase. Second, pre-processing is performed on the hand gesture data. The details of this phase are described in the data pre-processing subsection. Third, in the feature extractor phase, hyperparameter optimization is carried out, and then low- and high-level hand gesture features are obtained by utilising pre-trained networks. Fourth, in the fully connected layers (FC1, FC2, and FC3) phase, the training operation for the suggested architecture is performed using the optimised hyperparameter set obtained using GAs and hand gesture training datasets. Finally, the classification operations are performed using a softmax classifier in the output phase. The main contribution of the proposed framework is the introduction of a novel hybrid architecture that combines two widely used pre-trained network models in an optimised manner and uses a genetic algorithm for hyperparameter optimization. The proposed method was tested on three comprehensive datasets. The experimental results reveal that the method can effectively classify hand gestures with a high accuracy rate and that it outperforms the state-of-the-art methods in the literature. The performance results of the proposed method are explained in detail in the experimental results and discussions section.

The rest of this section is separated into three subsections: data pre-processing, an overview of the proposed model, and the literature used in the proposed method. First, in the data pre-processing subsection, hand gesture data in video format are represented as sequential single images. Second, the proposed hand gesture recognition framework is described in detail in the model overview subsection, and finally, in the utilised literature subsection, the hyperparameter search methods and deep learning architectures used in the proposed method are presented.

3.1 Data pre-processing

In this section, hand gesture samples are pre-processed for the proposed system. The pre-processing procedure, which is illustrated in Figure 1, comprises three phases. The samples in video format were first divided into image frames. Second, these frames were combined sequentially and transformed into sequential single images. Finally, sequential single images in RGB format were converted to grayscale since colours have no effect on classification and are only motion-oriented.

3.2 Overview of proposed model for hand gesture recognition

The proposed model is designed as a hybrid deep neural network, and it ensures an optimised framework for the task of hand gesture recognition. The suggested framework, which is shown in Figure 2, consists of five phases: gathering hand gesture data, pre-processing, designing the deep neural network architecture, optimising the hyperparameters, training, and evaluating the method. Additionally, Figure 3 shows a flowchart of the system, which illustrates these five steps in more detail. The pretrained networks in the pretraining and parameter tuning section are used for the task of feature extraction. The first three layers and the final layer in the training section represent fully connected (FC) layers that carry out learning operations and a softmax classifier that performs classification operations, respectively.

First, hand gesture samples are gathered from a personal dataset and the TinyHands [41] and LaRED [42] datasets. These hand gesture datasets are described in depth in the experimental results and discussion section. Then, the hand gesture data in video format are pre-processed. The details of this pre-processing phase are described in the previous subsection.

Next, the deep neural network architecture is designed. In this stage, the process of estimating an appropriate deep learning architecture is first performed. Pre-experiments uncovered that the hash model [43] results in better overall precision. For this reason, pre-trained architectures were combined inside a hash model. This hash module contains the GoogleNet and ResNet-50 architectures. Then, transfer learning was explored. Transfer learning is performed by discarding the pre-trained model's architecture up to a certain layer and then adding layers suitable for the current problem. Thus, the low- and medium-level feature extraction layers of the pre-trained model are transferred, and high-level feature extraction is performed with the layers that are added to the architecture later and selected in accordance with the relevant classification problem [44]. Transfer learning approaches have been extensively used to overcome various difficulties in classification operations. Examples of these difficulties are hardware resource constraints, model complexity, and time constraints. Therefore, transfer learning approaches were adapted for the suggested architecture to overcome these difficulties.

In the next step, the pre-processed hand gesture dataset is randomly divided into training and testing sets, with 70% of the dataset used for training and 30% used for testing. Then, 50% of the training set was used to create the development set. The development set [45], which is used for operations such as parameter tuning and feature extraction, is used in this study to adjust the initial learning rate, l_2 regularisation, and momentum parameters in the CNN training settings. Consequently, hyperparameter tuning is carried out using the development set derived from the training set and our deep learning-based architecture. Finally, the training stage is realised to reach a high accuracy rate.

To summarise, the proposed model based on transfer learning merges two pre-trained networks by applying an equal weighting process to generate a feature vector, and it uses a GA for hyperparameter optimization. Each phase is explained as follows: Initially, hand gesture data is collected and then pre-processed. Second, hyperparameter tuning is performed using the proposed deep learning-based architecture. Third, the pre-training process is performed. In this step, the GoogleNet and ResNet-50 architectures are trained on an ImageNet dataset [46]. Fourth, the features obtained from the GoogleNet and ResNet-50 architectures are merged to create a 4096-dimensional combined feature vector. At this point, he features produced by the GoogleNet and ResNet-50 architectures are extracted in the form of 2048-dimensional feature vectors from the final fully connected layers; the two architectures are shown in Figure 4 and Figure 5, respectively. Finally, the combined feature vector with 4096 elements is passed from the fully connected layers to the softmax layer so that it can be normalised. The fully connected layers comprise 4096 nodes and the softmax layer has 88 outputs, which correspond to 88 hand gestures.

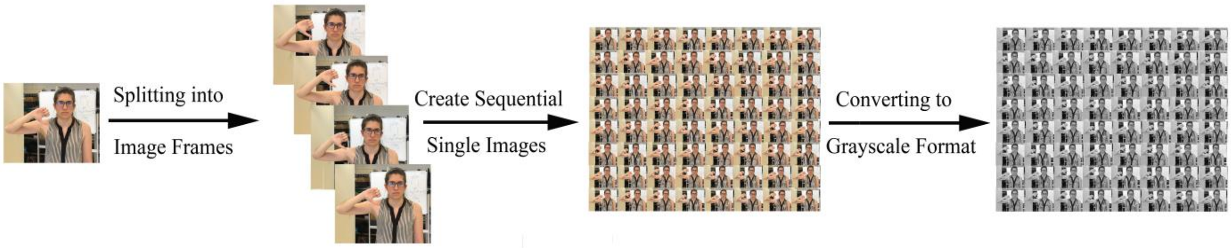


Figure 1. Pre-processing on a sample

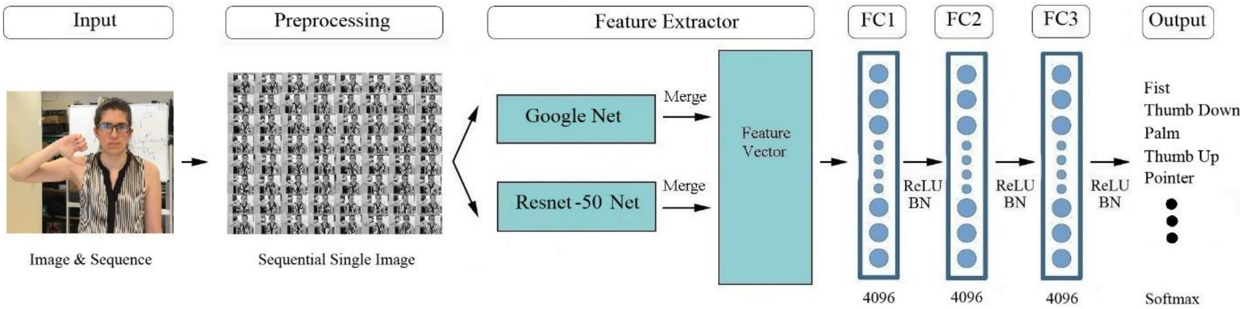


Figure 2. Proposed hand gesture recognition methodology

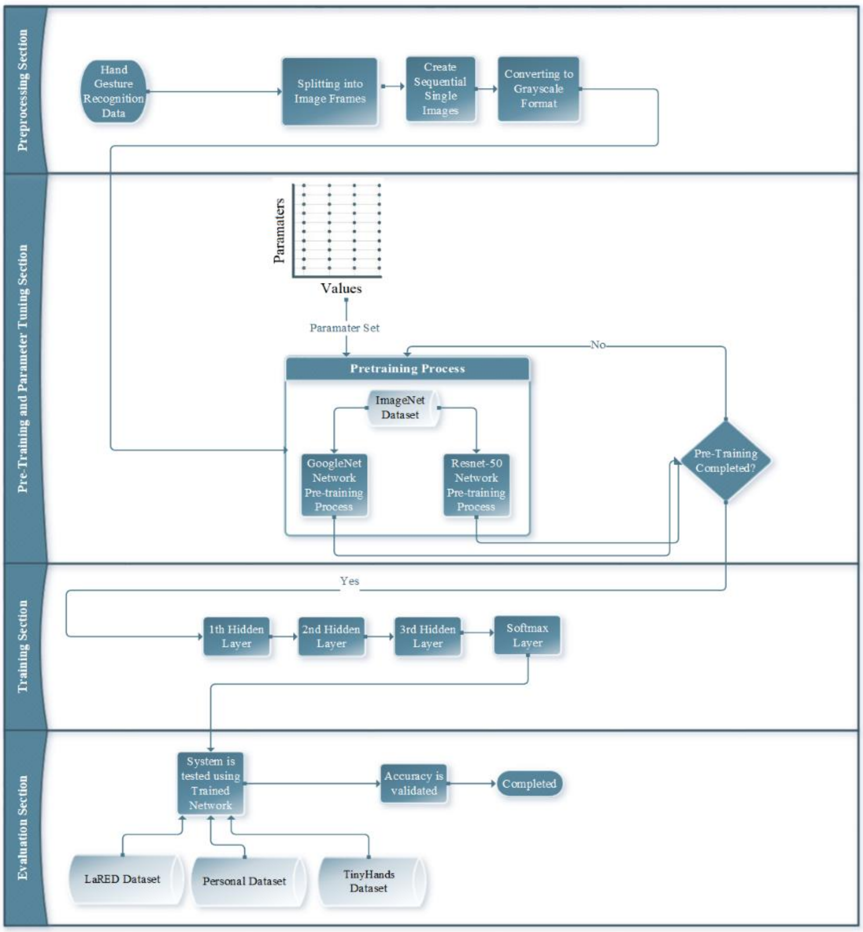


Figure 3. Flowchart of proposed deep learning architecture for hand gesture recognition

3.3 Literature used in the proposed method

This section reviews the literature that was used in the development of the proposed hand gesture architecture. Three hyperparameter search methods are discussed in detail in subsection 3.3.1, and two popular convolutional neural network architectures are discussed in detail in subsection 3.3.2.

3.3.1 Hyperparameter tuning methods

The grid search method (GSM) is a global search method used to optimise the parameters of a model. Trying to optimise all the parameters of a model with this method can be quite time consuming. Instead, it is possible to search for important parameters that will directly affect the results and seriously affect the model's performance. Possible different values are determined for each parameter and all combinations of these values are run for the relevant model. The parameter set that gives the best result according to the performance criteria is determined. The grid search method provides temporal benefits in the optimization process, and it ensures that the entire search space is searched at equal intervals. However, the risk of finding only the local best parameters is the disadvantage of the GSM [47].

The random search method (RSM) is a search method employed to optimise the parameters of a model, like the global search method. The lower and upper limits of the parameters to be optimised are determined. Then, the model is trained with a randomly determined parameter set for each epoch and the performance of the model is obtained. Temporal gains can be obtained with the RSM, which is performed with randomly determined parameter sets in the search space. Additionally, the probability of finding the global best set of parameters exists, although it is low. On the other hand, since the parameter sets are chosen randomly, the search may become stuck at certain points, and the search cannot be performed in every region of the search space [47].

A GA is a heuristic search algorithm proposed by John Holland in 1975 [48]. While chromosomes express general solutions, the main purpose of a GA is to find the gene sequence that gives the best fitness value in a solution space. Until the stopping criterion is met, the algorithm continues testing possible solutions and searching for the gene sequence that yields the most appropriate result.

3.3.2 Convolutional neural network architectures

The GoogleNet network, shown in Figure 4, was the first convolutional neural network to use the notion of width. It employs the Inception module, which is comprised of deeper branches and a shortcut branch used to obtain the width item in the model. This module also allows the network to choose between multiple convolutional filter sizes in each block. The Inception module, which is presented in Figure 4, uses 3×3 maximum pooling and 5×5, 3×3, and 1×1 filters in the convolution layers in a parallel manner [49]. The idea behind this is that convolution filters of different sizes will handle objects at multiple scales better. The GoogleNet architecture is 22 layers deep, with 27 pooling layers. The architecture, as shown in Figure 4, includes four main sections: a stacked Inception module, auxiliary classifiers, an output classifier, and a stem. This architecture includes 144 layers, including an output layer, convolutional layers, maxpooling layers, a softmax layer, an input layer, rectified linear unit (ReLU) layers, and fully connected layers. Furthermore, GoogleNet employs nine Inception modules and around 7 million parameters [49].

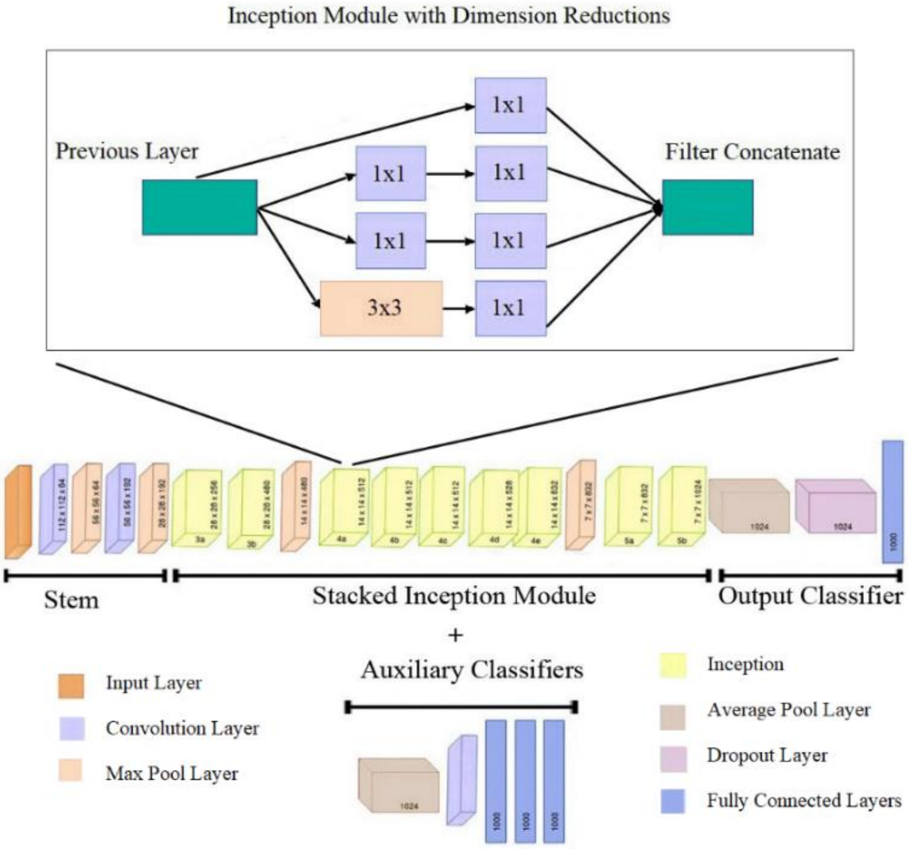


Figure 4. GoogleNet network [49]

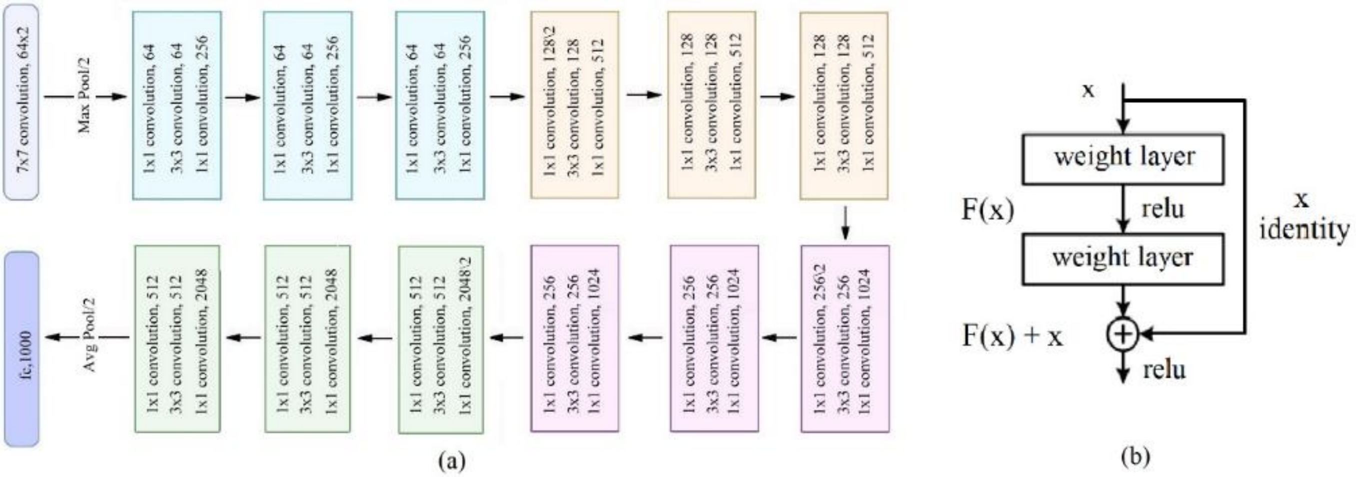


Figure 5. ResNet-50 network (a) and residual block (b) [50]

The ResNet-50 network [50], proposed by He et al. from Microsoft Research Asia in 2015, is a convolutional neural network that is 50 layers deep. The ResNet model aims to solve the performance degradation problem of CNNs. It adds shortcuts between layers to solve this problem. This structure, called a residual learning block, is shown in Figure 5(b). Residual learning blocks are considered the building blocks of ResNet. In residual blocks, the input x is added directly to the output of the network, i.e., F(x) + x, and this path is known as a jump link or shortcut. The ResNet-50 architecture, illustrated in Figure 5(a), includes a fully connected layer, two pooling operations, five convolutional blocks, and a softmax layer, and it comprises 25.6 million parameters.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the datasets used, experimental results, an assessment of the proposed model, and implementation details. The implementation of the proposed methodology was performed using the Python scripting language. To perform our experiments, we used a personal computer with an AMD Ryzen 5 5600H processor with a speed of 4.2 GHz and 64 GB of RAM. Additionally, our experiments were conducted in a Linux environment.

Table 1. Parameter values used in GSM

Parameters	Values
Initial learning rate	0.001, 0.005, 0.01
l_2 regularisation	0.00001, 0.0001, 0.001
Momentum	40, 60, 50

In the proposed method, the GSM, RSM, and GA are used for parameter adjustment. The initial learning rate (ILR), l_2R regularisation (l_2R), and momentum (M) parameter values used for the GSM are shown in Table 1. The lower limits of the initial learning rate, l_2 regularisation, and momentum values for the RSM are 0.001, 0.00001, and 0.75, respectively, while the upper limits are 0.01, 0.001, and 0.95, respectively. The lower and upper limits used for the RSM are also used for the GA. In the GA, the mutation rate is set to 20, the crossover rate is set to 0.80, and the iteration number is set to 40. The training parameters of the proposed model have the following values: the minibatch size is 64 and the maximum number of epochs is 120. The stochastic gradient descent method with momentum was used in the training of the model. In addition, training and test samples were chosen randomly from among the datasets used, and evaluation operations were carried out individually on the proposed method. Seventy percent of the available data was employed to train our proposed model, whereas the remaining 30% was used to test the model.

4.1 Benchmark datasets

Our experiments were realised using three exhaustive datasets, namely, the TinyHands and LaRED datasets and a personal dataset. The characteristics of these three datasets are summarised in Table 2, and sample images from each dataset are illustrated in Figure 6. Additionally, these three datasets are described in detail below.

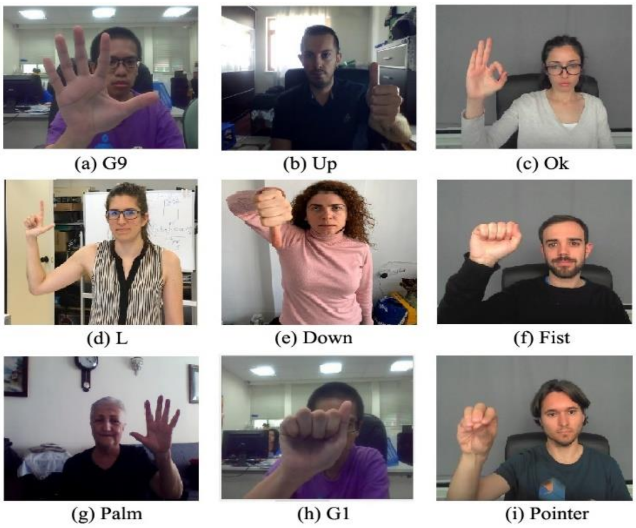

(a) G9
(b) Up
(c) Ok
(d) L
(e) Down
(f) Fist
(g) Palm
(h) G1
(i) Pointer

Figure 6. Sample images from hand gesture recognition datasets [41, 42]

The LaRED dataset [42] is a large RGB-D hand gesture dataset containing approximately 240,000 tuples of images. Each tuple includes a depth image, a mask of the hand region, and a colour image. This dataset is one of the largest hand movement datasets in the literature, with 81 different classes; it was created by acquiring 27 hand movements (denoted as $\boldsymbol{G_1}, \boldsymbol{G_2}, \boldsymbol{G_3}, \dots, \boldsymbol{G_{27}}$) in three different rotations. The dataset was created with a short-range Intel depth camera. Most of the

27 hand gesture classes are from American Sign Language; the remainder consist of some common gestures used for interacting with electronic devices. There were 10 subjects: five were female and five were male. Three hundred motion images per class were obtained from each subject by having them repeat the same hand gesture with small motions. There are approximately 218,700 training set samples and 24,300 test set samples in this dataset. In our study, only the colour images of VGA resolution of this dataset were employed, and the rest of the dataset was not included in our study.

The TinyHands dataset [41] was created in two different installation environments: one had a simple background and the other had a complex background. Half of the data was obtained in a simple environment and the other half was obtained in a complex environment. In the dataset, the complex background has various illumination conditions and a highly cluttered environment. In the dataset, hand movements are performed at different positions, and the hands only cover about 10% of the field of the entire image. Forty subjects were used to create this dataset and approximately 1,400 hand gesture images were obtained for each hand gesture class by having each subject perform seven different hand gestures (e.g., fist, L, ok, pointer, up). There are approximately 294,000 training set samples and 98,000 test set samples in the dataset.

The personal dataset was obtained by the author of the present paper with 10 subjects. There are 7 different hand gesture classes in the created dataset. These classes are ‘fist’, ‘L’, ‘arrow’, ‘palm’, ‘pointer’, ‘down’, and ‘up’. In the dataset, 400 motion images per hand gesture class were acquired from each subject by having them repeat the same hand gesture with small motions. There are approximately 25,200 training set samples and 2,800 test set samples in this dataset.

4.2 Results and discussion

The GSM, RSM, and GA are used on the LaRED dataset to optimise the training parameters of the proposed model. The random parameter sets for the RSM and GSM and the validation errors of these parameter sets are presented in Table 3. According to Table 3, the parameter values that give a validation error of 0.3106 for the GSM and 0.3113 for the RSM are the best parameter values obtained as a result of the GSM and the RSM. When the GA is run for parameter optimization, the parameter set that gives a validation error (VE) of 0.3118 was given as output. The following parameter values were obtained: ILR = 0.0026, 1_2 R = 0.00096, and M = 0.85. The results obtained on the LaRED dataset with the hyperparameter methods used are shown in Table 5. According to this table, the best accuracy result was obtained with the GA; therefore, this method was used in our study to determine the training parameters of the proposed deep neural network architecture.

Table 2. Characteristics of datasets used for hand gesture recognition

Dataset	Number of Subjects	Number of Hand Gestures	Training Set	Testing Set
LaRED	10	81	~218700	~24300
TinyHands	40	7	~294000	~98000
Personal	10	7	~25200	~2800

Table 3. Parameter values and validation errors determined by the GSM and RSM

GSM				RSM			
ILR	l_2R	**M**	**VE**	**ILR**	l_2R	**M**	**VE**
0.001	0.00001	0.75	0.3588	0.0095	0.00003	0.77	0.3481
0.001	0.00001	0.85	0.3623	0.0080	0.00025	0.76	0.3864
0.001	0.00001	0.95	0.3124	0,0063	0.00072	0.81	0.3628
0.001	0.00010	0.75	0.3876	0.0042	0.00049	0.95	0.4695
0.001	0.00010	0.85	0.3788	0.0051	0.00020	0.84	0.3766
0.001	0.00010	0.95	0.3277	0.0072	0.00048	0.78	0.3935
0.001	0.00100	0.75	0.3855	0.0075	0.00078	0.75	0.3382
0.001	0.00100	0.85	0.3672	0.0067	0.00010	0.85	0.3529
0.001	0.00100	0.95	0.3699	0.0032	0.00006	0.92	0.3774
0.005	0.00001	0.75	0.3702	0.0011	0.00029	0.85	0.3946
0.005	0.00001	0.85	0.3866	0.0055	0.00068	0.92	0.4234
0.005	0.00001	0.95	0.4352	0.0093	0.00045	0.88	0.4086
0.005	0.00010	0.75	0.3547	0.0064	0.00061	0.95	0.5271
0.005	0.00010	0.85	0.3602	0.0090	0.00045	0.92	0.9365
0.005	0.00010	0.95	0.4718	0.0018	0.00066	0.78	0.3544
0.005	0.00100	0.75	0.3766	0.0097	0.00026	0.76	0.3862
0.005	0.00100	0.85	0.3106	0.0057	0.00030	0.84	0.3628
0.005	0.00100	0.95	0.4386	0.0059	0.00069	0.78	0.3113
0.010	0.00001	0.75	0.3765	0.0021	0.00048	0.89	0.3288
0.010	0.00001	0.85	0.3844	0.0091	0.00048	0.85	0.4483
0.010	0.00001	0.95	0.8988	0.0100	0.00054	0.83	0.3786
0.010	0.00010	0.75	0.3418	0.0042	0.00054	0.85	0.4348
0.010	0.00010	0.85	0.3879	0.0085	0.00006	0.88	0.4456
0.010	0.00010	0.95	0.9131	0.0078	0.00011	0.76	0.3620
0.010	0.00100	0.75	0.3286	0.0044	0.00010	0.92	0.3765
0.010	0.00100	0.85	0.3982	0.0078	0.00078	0.85	0.3998
0.010	0.00100	0.95	0.9255	0.0012	0.00058	0.95	0.3435

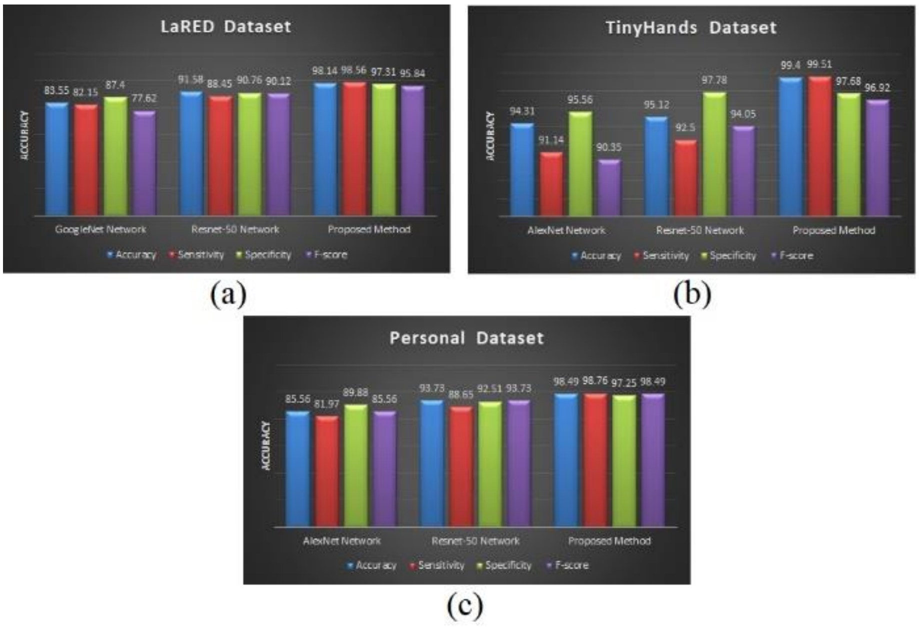
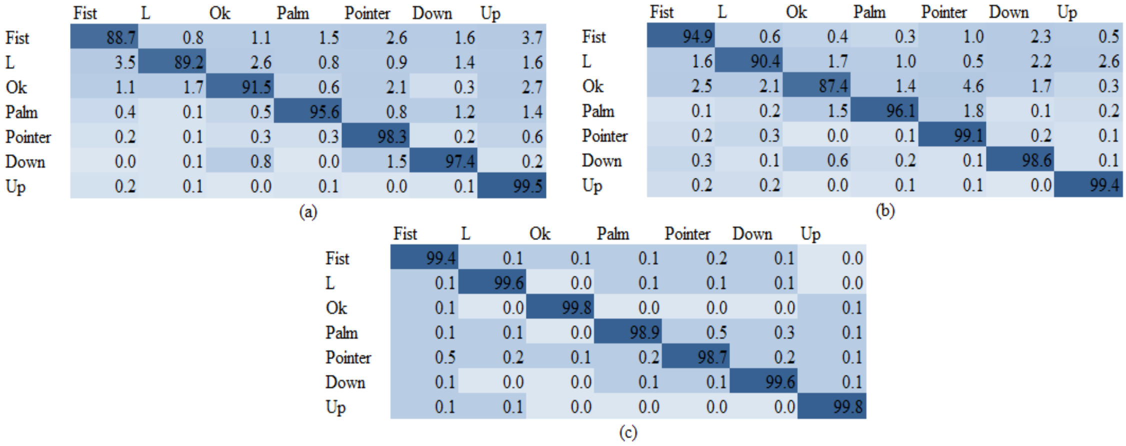


Figure 7. Quantitative results for the LaRED (a), TinyHands (b), and personal (c) datasets



	Fist	L	Ok	Palm	Pointer	Down	Up
Fist	88.7	0.8	1.1	1.5	2.6	1.6	3.7
L	3.5	89.2	2.6	0.8	0.9	1.4	1.6
Ok	1.1	1.7	91.5	0.6	2.1	0.3	2.7
Palm	0.4	0.1	0.5	95.6	0.8	1.2	1.4
Pointer	0.2	0.1	0.3	0.3	98.3	0.2	0.6
Down	0.0	0.1	0.8	0.0	1.5	97.4	0.2
Up	0.2	0.1	0.0	0.1	0.0	0.1	99.5

(a)

	Fist	L	Ok	Palm	Pointer	Down	Up
Fist	94.9	0.6	0.4	0.3	1.0	2.3	0.5
L	1.6	90.4	1.7	1.0	0.5	2.2	2.6
Ok	2.5	2.1	87.4	1.4	4.6	1.7	0.3
Palm	0.1	0.2	1.5	96.1	1.8	0.1	0.2
Pointer	0.2	0.3	0.0	0.1	99.1	0.2	0.1
Down	0.3	0.1	0.6	0.2	0.1	98.6	0.1
Up	0.2	0.2	0.0	0.1	0.1	0.0	99.4

(b)

	Fist	L	Ok	Palm	Pointer	Down	Up
Fist	99.4	0.1	0.1	0.1	0.2	0.1	0.0
L	0.1	99.6	0.0	0.1	0.1	0.1	0.0
Ok	0.1	0.0	99.8	0.0	0.0	0.0	0.1
Palm	0.1	0.1	0.0	98.9	0.5	0.3	0.1
Pointer	0.5	0.2	0.1	0.2	98.7	0.2	0.1
Down	0.1	0.0	0.0	0.1	0.1	99.6	0.1
Up	0.1	0.1	0.0	0.0	0.0	0.0	99.8

(c)

Figure 8. Confusion matrices for the TinyHands dataset for seven hand gestures for GoogleNet (a), ResNet-50 (b), and the proposed network (c)

Table 4. Formulas used to calculate the evaluation metrics

Metric	Formula
F-score	2*TP / (2*TP+FP+FN)
Specificity	TN / (TN+FP)
Sensitivity	TP / (TP+FN)
Accuracy	(TP+TN) / (TP+TN+FP+FN)

Table 5. Comparison with existing state-of-the-art algorithms using the LaRED dataset

Model/Method	*Accuracy (%)*
Sanchez-Riera et al. [51] / Deep belief networks	66.13
Sanchez-Riera et al. [51] / Restricted Boltzmann machines	72.95
Hsiao et al. [42] / Baseline network	74.55
Sanchez-Riera et al. [51] / Stacked autoencoders	81.09
Sanchez-Riera et al. [51] / Convolutional neural network	88.72
Mohammed et al. [52] / Lightweight MobileNet	97.25
GS + proposed method	93.56
RS + proposed method	88.92
GA + proposed method	**98.14**

Assessment metrics used for classification processes are crucial for figuring out the yield and performance of deep neural network models. Assessment metrics explain the performance of the classification model and discriminate among model results [53]. The sensitivity, F-score, specificity, and accuracy metrics were therefore utilised to indicate the classification performance of the suggested method. These assessment metrics were computed using the equations shown in Table 4. In this table, FN stands for false negative, TN stands for true negative, FP stands for false positive, and TP stands for true positive.

These performance metrics are the first step in evaluating the performance of the proposed deep neural network architecture. Comparisons of GoogleNet and ResNet-50 networks with the proposed model are performed. The metric values obtained from GoogleNet, ResNet-50, and the proposed deep neural network architecture for the LaRED, TinyHands, and personal datasets are presented in Figure 7. According to this figure, our proposed model performs better than the GoogleNet and ResNet-50 networks. In addition, the results from the proposed model indicate similar performance outcomes on the LaRED, TinyHands, and personal datasets, whereas the performance results of the GoogleNet and ResNet-50 networks vary considerably between the LaRED, TinyHands, and personal datasets. The proposed model has the best performance out of all the advanced models compared in our study. Second, hand gesture types were examined using confusion matrices. Figure 8 presents the confusion matrices for the TinyHands dataset for seven hand gesture types (fist, L, ok, palm, pointer, down, up) obtained from the proposed, ResNet-50, and GoogleNet deep neural network architectures. Here, the accuracy values of seven hand gesture types are shown by confusion matrices. According to Figure 8(c), the suggested method yields good accuracy values for whole hand gesture classifications, apart from the 'pointer' hand gesture type. The three network models can easily distinguish the 'up' hand gesture type, and the ResNet-50 model, as illustrated in Figure 8(b), detects the 'pointer' hand gesture type better than the other models. Finally, a comparison with other state-of-the-art hand recognition methods was performed to evaluate the performance of the proposed model. Table 5 shows the accuracy values obtained on the LaRED dataset by state-of-the-art models and the proposed network. The proposed model is more efficient and accurate than other related techniques.

5. CONCLUSION

Hand gesture recognition is a very important and challenging task in many interactive applications, such as driver hand behaviour monitoring, hand action analysis, virtual reality, etc. In recent years, several algorithms have been developed to solve this problem. However, a robust and efficient system has still not been created due to challenging conditions such as hardware resource constraints, model complexity, extreme dataset dimensions, etc. Therefore, our study suggests a new hash deep learning architecture for recognising hand gestures effectively using the transfer learning method. Initially, hand gesture data were gathered from assorted exhaustive datasets. Second, the hand gesture data were pre-processed. Third, hyperparameter adjustment was performed with the development set derived from the training set and our deep learning-based architecture. In this step, a GA was used for parameter tuning because it gave the best accuracy rate. Fourth, the features were extracted using pre-trained networks. Lastly, the training was performed using the optimised hyperparameter set obtained using GAs and hand gesture training datasets.

The study makes two main contributions to the literature on hand gesture recognition. The first is the creation of a new database for hand gesture recognition. The second is a novel hash architecture that combines two widely used pre-trained network models in an optimised manner and uses a genetic algorithm for hyperparameter optimization. Experiments were performed using benchmark datasets to show the effectiveness and robustness of the proposed approach. In these experiments, the suggested architecture was first compared with each individual model. The proposed architecture outperforms individual models in many scenarios according to the sensitivity, F-score, specificity, and accuracy metrics. Next, the suggested architecture was compared with state-of-the-art models. The obtained results also reveal that the suggested method is superior to state-of-the-art methods. For future studies, we plan to use different global and metaheuristic search methods for parameter optimization, and we plan to compare the performances of more models. Additionally, transfer learning can be performed and a performance analysis can be made using different pre-trained models, i.e., models other than ResNet-50 and GoogleNet. In addition, we also plan to test the proposed method with different datasets.

REFERENCES

[1] World Health Organization. WHO - Deafness and Hearing Loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, accessed on March 1, 2022.

[2] Rautaray, S.S., Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: A survey. Artificial Intelligence Review, 43(1): 1-54. https://doi.org/10.1007/s10462-012-9356-9

[3] Hall, E.T. (1973). The Silent Language. Anchor Books.

[4] Karam, M. (2006). A framework for research and design of gesture-based human-computer interactions. PhD dissertation. Department of Electronics and Computer Science, University of Southampton, Southampton, UK.

[5] Molchanov, P., Gupta, S., Kim, K., Kautz, J. (2015). Hand gesture recognition with 3D convolutional neural networks. In 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, pp. 1-7. https://doi.org/10.1109/CVPRW.2015.7301342

[6] Lyer, P., Tarekar, S., Dixit, S. (2019). Hand gesture controlled robot. In 2019 9th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-19), Nagpur, India, pp. 1-5. https://doi.org/10.1109/ICETET-SIP-1946815.2019.9092032

[7] Kang, T., Chae, M., Seo, E., Kim, M., Kim, J. (2020). DeepHandsVR: Hand interface using deep learning in immersive virtual reality. Electronics, 9(11): 1863. https://doi.org/10.3390/electronics9111863

[8] Arathi, P.N., Arthika, S., Ponmithra, S., Srinivasan, K., Rukkumani, V. (2017). Gesture based home automation system. In 2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2), Chennai, India, pp. 198-201. https://doi.org/10.1109/ICNETS2.2017.8067929

[9] Kumar, P., Verma, J., Prasad, S. (2012). Hand data glove: A wearable real-time device for human computer interaction. International Journal of Advanced Science and Technology, 43(6): 15-26.

[10] Schwenk, H. (2012). Continuous space translation models for phrase-based statistical machine translation. In Proceedings of COLING 2012: Posters, pp. 1071-1080.

[11] Chatfield, K., Arandjelović, R., Parkhi, O., Zisserman, A. (2015). On-the-fly learning for visual search of large-scale image and video datasets. International Journal of Multimedia Information Retrieval, 4(2): 75-93. https://doi.org/10.1007/s13735-015-0077-0

[12] Pi, Y., Nath, N.D., Behzadan, A.H. (2020). Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. Advanced Engineering Informatics; 43: 101009. https://doi.org/10.1016/j.aei.2019.101009

[13] Gu, B., Ge, R., Chen, Y., Luo, L., Coatrieux, G. (2021). Automatic and robust object detection in x-ray baggage inspection using deep convolutional neural networks.

IEEE Transactions on Industrial Electronics, 68(43): 10248-10257. https://doi.org/10.1109/TIE.2020.3026285

[14] Maas, A.L., Qi, P., Xie, Z., Hannun, A.Y., Lengerich, C.T., Jurafsky, D., Ng A.Y. (2017). Building DNN acoustic models for large vocabulary speech recognition. Computer Speech and Language, 41: 195-213. https://doi.org/10.1016/j.csl.2016.06.007

[15] Yilmaz, A.A., Guzel, M.S., Bostanci, E., Askerzade, I. (2020). A novel action recognition framework based on deep-learning and genetic algorithms. IEEE Access, 8: 100631-100644. https://doi.org/10.1109/ACCESS.2020.2997962

[16] Aslan, Ö., Yilmaz, A.A. (2021). A new malware classification framework based on deep learning algorithms. IEEE Access, 9: 87936-87951. https://doi.org/10.1109/ACCESS.2021.3089586

[17] Yilmaz, A.A., Guzel, M.S., Askerzade I., Bostanci, E. (2018). A vehicle detection approach using deep learning methodologies. In Proceedings of International Conference on Theoretical and Applied Computer Science and Engineering (ICTACSE), İstanbul, Turkey, pp. 64-71. https://doi.org/10.48550/arXiv.1804.00429

[18] Sille, R., Choudhury, T., Chauhan, P., Sharma, D. (2021). Dense hierarchical CNN – A unified approach for brain tumor segmentation. Revue d'Intelligence Artificielle, 35(3): 223-233. https://doi.org/10.18280/ria.350306

[19] Li, F., Tran, L., Thung, K.H., Ji, S., Shen, D., Li, J. (2015). A robust deep model for improved classification of AD/MCI patients. IEEE Journal of Biomedical and Health Informatics, 19(5): 1610-1616. https://doi.org/10.1109/JBHI.2015.2429556

[20] Yılmaz, A.A., Guzel, M.S., Askerbeyli, I., Bostancı, E., Özseven, T. (2020). A hybrid facial emotion recognition framework using deep learning methodologies. In Human-Computer Interaction. Nova Science Publishers.

[21] Cao, N.D., Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs. arXiv preprint arXiv: 11973. https://doi.org/10.48550/arXiv.1805.11973

[22] Xie, T., Grossman, J.C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters, 120(14): 145301. https://doi.org/10.1103/PhysRevLett.120.145301

[23] Maraqa, M., Abu-Zaiter, R. (2008). Recognition of Arabic Sign Language (ArSL) using recurrent neural networks. In 2008 First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT), pp. 478-481. https://doi.org/10.1109/ICADIWT.2008.4664396

[24] Flores, C.J.L., Cutipa, A.G., Enciso, R.L. (2017). Application of convolutional neural networks for static hand gestures recognition under different invariant features. In 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON). pp. 1-4. https://doi.org/10.1109/INTERCON.2017.8079727

[25] Alashhab, S., Gallego, A.J., Lozano, M.Á. (2018). Hand gesture detection with convolutional neural Networks. In 2018 International Symposium on Distributed Computing and Artificial Intelligence (DCAI), Salamanca, Spain, pp. 45-52. https://doi.org/10.1007/978-3-319-94649-8_6

[26] Pigou, L., Dieleman, S., Kindermans, P.J., Schrauwen, B. (2014). Sign language recognition using convolutional neural networks. In European Conference on Computer Vision, pp. 572-578. https://doi.org/10.1007/978-3-319-16178-5_40

[27] Rao, G., Syamala, K., Kishore, P., Sastry, A. (2018). Deep convolutional neural networks for sign language recognition. In 2018 Conference on Signal Processing and Communication Engineering Systems (SPACES), pp. 194-197. https://doi.org/10.1109/SPACES.2018.8316344

[28] Jiang, L., Xia, M., Liu, X., Bai, F. (2020). Givs: Fine-grained gesture control for mobile devices in driving environments. IEEE Access, 8: 49229-49243. https://doi.org/10.1109/ACCESS.2020.2971849

[29] Baig, F., Fahad Khan, M., Beg, S. (2013). Text writing in the air. Journal of Information Display, 14(4): 137-148. https://doi.org/10.1080/15980316.2013.860928

[30] Höll, M., Oberweger, M., Arth, C., Lepetit, V. (2018). Efficient physics-based implementation for realistic hand-object interaction in virtual reality. 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, pp. 175-182. https://doi.org/10.1109/VR.2018.8448284

[31] Shen, Y., Ong, S.K., Nee, A.Y.C. (2011). Vision-Based hand interaction in augmented reality environment. International Journal of Human–Computer Interaction, 27(6): 523-544. https://doi.org/10.1080/10447318.2011.555297

[32] Rao, J., Gao. T., Gong, Z., Jiang, Z. (2009). Low cost hand gesture learning and recognition system based on hidden Markov model. In Proceedings of the 2009 Second International Symposium on Information Science and Engineering (ISISE), pp 433-438. https://doi.org/10.1109/ISISE.2009.53

[33] Gupta, S., Jaafar, J., Ahmad, W.F.W. (2012). Static hand gesture recognition using local Gabor filter. Procedia Engineering, 41: 827-832. https://doi.org/10.1016/j.proeng.2012.07.250

[34] Rahman, M.H., Afrin, J. (2013). Hand gesture recognition using multiclass support vector machine. International Journal of Computer Applications, 74(1): 39-43. https://doi.org/doi: 10.5120/12852-9367

[35] Marium, A., Rao, D., Crasta, D.R., Acharya, K., D'Souza, R. (2017). Hand Gesture Recognition using Webcam. American Journal of Intelligent Systems, 7(3): 90-94. https://doi.org/10.5923/j.ajis.20170703.11

[36] Oyedotun, O.K., Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition. Neural Computing and Applications, 28(12): 3941-3951. https://doi.org/10.1007/s00521-016-2294-8

[37] Thomas Moeslund's gesture recognition database. PRIMA. http://www-prima.inrialpes.fr/FGnet/data/12-MoeslundGesture/database.html, accessed on March 1, 2022.

[38] Devineau, G., Moutarde, F., Xi, W., Yang, J. (2018). Deep learning for hand gesture recognition on skeletal data. In 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), Xi'an, China, pp. 106-113. https://doi.org/10.1109/FG.2018.00025

[39] Chevtchenko, S.F., Vale, R.F., Macario, V., Cordeiro, F.R. (2018). A convolutional neural network with feature fusion for real-time hand posture recognition. Applied

Soft Computing, 73: 748-766. https://doi.org/10.1016/j.asoc.2018.09.010

[40] Wu, D., Pigou, L., Kindermans, P.J., Le, N.D.H., Shao, L., Dambre, J., Odobez, J.M. (2016). Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8): 1583-1597. https://doi.org/10.1109/TPAMI.2016.2537340

[41] Bao, P., Maqueda, A.I., Del-Blanco, C.R., Garciá, N. (2017). Tiny hand gesture recognition without localization via a deep convolutional network. IEEE Transactions on Consumer Electronics, 63(3): 251-257. https://doi.org/10.1109/TCE.2017.014971

[42] Hsiao, Y.S., Sanchez-Riera J., Lim, T., Hua K.L., Cheng, W.H. (2014). LaRED: A large RGB-D extensible hand gesture dataset. In 2014 5th ACM Multimedia Systems Conference (MMSys'14), pp. 53-58. https://doi.org/10.1145/2557642.2563669

[43] Alonso, D.G., Teyseyre, A., Berdun, L., Schiaffino, S. (2019). A deep learning approach for hybrid hand gesture recognition. In Mexican International Conference on Artificial Intelligence, pp. 87-99. https://doi.org/10.1007/978-3-030-33749-0_8

[44] Ozcan, T., Basturk, A. (2019). Lip reading using convolutional neural networks with and without pre-trained models. Balkan Journal of Electrical and Computer Engineering, 7(2): 195-201. https://doi.org/10.17694/bajece.479891

[45] Ng, A. (2018). Machine learning yearning: Technical strategy for AI engineers, in the era of deep learning. Deeplearning.AI, USA.

[46] The ImageNet Dataset, http://www.image-net.org/, accessed on March 1, 2022.

[47] Online courses and lessons about data science, machine learning and artificial intelligence. YOUR DATA TEACHER.https://www.yourdatateacher.com/2021/05/19/hyperparameter-tuning-grid-search-and-random-search/, accessed on March 1, 2022.

[48] Holland, J.H. (1992). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press. https://doi.org/10.7551/mitpress/1090.001.0001

[49] Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[50] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[51] Sanchez-Riera, J., Hua, K.L., Hsiao, Y.S., Lim, T., Hidayati, S.C., Cheng, W.H. (2016). A comparative study of data fusion for RGB-D based visual recognition. Pattern Recognition Letters, 73: 1-6. https://doi.org/10.1016/j.patrec.2015.12.006

[52] Mohammed, A.A.Q., Lv, J., Islam, M.S. (2019). A deep learning-based end-to-end composite system for hand detection and gesture recognition. Sensors, 19(23): 5282. https://doi.org/10.3390/s19235282

[53] AnalyticsVidhya, https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics, accessed on March 1, 2022.