International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Face and Hand Gesture Recognition Based Person Identification System using Convolutional Neural Network

Mysha Sarin Kabisha[1], Kazi Anisa Rahim[2], Md. Khaliluzzaman*[3], Shahidul Islam Khan[4]

Submitted: 21/11/2021 Accepted : 07/02/2022

Abstract: Person identification system is now become the most hyped system for security purpose. It's also gaining a lot of attention in the field of computer vision. For verification of human, facial recognition and hand gesture recognition are the most common topics of research. In the current days, various researchers focused on facial and hand gesture recognition using various shallow techniques and Deep Convolutional Neural Network (DCNN). However, using one feature of human for person identification is the most researched topic till now. In this paper, we proposed a Convolutional Neural Network (CNN) based system which will identify a person using two traits i.e., face and hand gesture of number sign of that person. For feature extraction and recognition Neural Network have shown immense good result. This proposed system works on two models, one is a VGG16 architecture model for face recognition and another model is for hand gesture which is based on simple CNN with two convolutional layers. With two customized dataset our face model gained 98.00% accuracy and hand gesture (number sign) model gained an accuracy of 98.33%.

Keywords: *Deep Learning, Convolutional Neural Network, Face Recognition, Hand Gesture Recognition, VGG16*

1. Introduction

Security refers to the measures that mainly taken to be safe or protected from the any kind of harm. The term security system is the consequence of ensuring security through a system for all. It is literally a mean by which something is secured through a system. Different kinds of security systems are available now a days, and people are researching a lot in this field. Security through Person Identification or Verification is one of the popular approaches among all. Person identification can be done using different biometric process like Iris recognition, fingerprint recognition, face recognition, and gait. In recent years, person identification has been gaining great attention, and many works have done for successful identification of person. Identifying a person in real time is a system that allows access to a portion of people while denying access to others. Person recognition can also performed by using facial features along with hand gesture based. This operation can be performed through the Deep Learning (DL) approach. Deep Learning is a type of machine learning that use numerous layers to extract higher-level features from unprocessed data such as images and videos. Artificial Neural Networks (ANN) is the foundation of modern deep learning approaches. Many Deep Learning-based facial and gesture recognition algorithms have recently shown promising results with large-scale labeled examples [6].

Face recognition and hand gesture recognition are the two most common research areas of person identification. Face recognition is a method for identifying individuals using their face [4]. Due to the rapid growth of Artificial Intelligence (AI), face recognition is again in talk. From a lot of previous research on face recognition we came to know that traditional shallow learning-based methods have been facing a lot of difficulties like facial disguise, pose variation, and complexity of the background. So now, Deep Learning-based methods are taking its place in image processing because of its ability of extracting more features than shallow models [2]. Beside face recognition, for human-computer interaction, hand gesture recognition is vital. Gestures can be derived from any physical move or mood; however, they are most typically associated with the face and hand. Computer vision has been used in a variety of ways to recognize hand gestures. Hand gesture recognition is critical in a variety of applications, including sign language for the disabled, recognition of sign languages, robot control, and virtual reality.

It is a simple task for a human to analyze and recognize a large number of details in a visual scene. For a computer, however, this is a difficult task that necessitates a large amount of calculation and memory [3]. Different CNN like 2D CNN, 3D CNN, Deep Neural Networks have also been using for recognition approaches. Machine Learning is a self-adaptive, which performs on the data that are given to it and shows progress over the time with experience or newly added data. Machine learning needs some guidance to work, which might not be so appropriate for the fast-running world. So, DL comes to rescue of machine learning. Deep Learning carries out the process of machine learning by utilizing a hierarchical level of ANN. There are different Deep Learning approaches available like Autoencoders, Deep Belief Net, CNN, and RNN. The ANNs are built by imitating human brains where neuron nodes connected with each other like a web. Deep networks have the ability of learning from unstructured or

[1] *Dept. of CSE, International Islamic University Chittagong, Chattogram-4318, Bangladesh. ORCID ID:0000-0003-1551-9908*
[2] *Dept. of CSE, International Islamic University Chittagong, Chattogram-4318, Bangladesh. ORCID ID:0000-0001-9224-566X*
[3] *Dept. of CSE, International Islamic University Chittagong, Chattogram-4318, Bangladesh. ORCID ID:0000-0001-6846-1610*
[4] *Dept. of CSE, International Islamic University Chittagong, Chattogram-4318, Bangladesh. ORCID ID:0000-0002-8740-2744*
** Corresponding Author Email: khalil@iiuc.ac.bd*

unlabeled raw data. It works like human brain where "Neuron" indicates a mathematical function that collects and classifies information according to specific architecture. In neural network each node in the layer of interconnected nodes is a perceptron. The signal produced by multiple linear regressions into an activation function is feeding by the perceptron. Neural network mainly comprised of three layers. These are, Input Layer: Takes primary data through corresponding layers for further analysis, then Hidden Layers: It's the intermediate layer where every computation is done and activation function provides the output and Output Layer: It's the last layer that brings out the information learned by the network.

With the fast development of deep neural networks, CNN is showing better results in deep feature learning. Comparing to other previous methods, it's more efficient and faster in learning important features without human supervision. And there is no need of extra feature extraction step. For detection and recognition model CNN is proving itself as a better one than other shallow models.

In the current time, security becomes an important issue of daily life. Many different security systems are presently used. One of them is surveillance camera, which helps to recognize the person. Facial recognition, which is another recognition system, mostly used for security purpose. For making the security system more secure, in this paper introduces dual feature of human simultaneously for person identification. Here, for person identification, firstly, person face is recognized. After confirmation the person, the hand gesture of that person is recognized. If the both face and hand gesture is verified the person is recognized as authentic. Otherwise, the person is threaded as an unauthentic person.

The body of this work is divided into the following sections. The associated literature is described in Section 2. The methodology of the proposed security system is presented in Section 3. The experimental results, as well as datasets, are presented in Section 4. Section 5 brings the paper to a conclusion.

2. Related Work

In literature, we have found a good number of research-works on Face Recognition and Hand Gesture Recognition separately. These section summaries the Face Recognition and Hand Gesture recognition methods that have been implemented and tested previously. In [1], suggested a model for extracting the unique properties of every human facial image for identification and recognition using a feature extraction module. They also employed LBP to improve the performance of the facial recognition system. Eigen face technique is utilized for face acknowledgement. Though they acquired a good performance for surveillance but their method is based on some complex method.

A CNN based architecture is developed for Face recognition in [2]. Batch normalization process used after every two different layers in training stage and softmax classifier for classification stage. Gained satisfying accuracy by testing this model with Georgia Tech Fest Database and tried different scenarios by changing image size, batch size, learning rate etc. In [3], introduced a CNN based architecture using two convolution layers for face recognition and got 98.7% accuracy on ORL dataset, which is built by the AT&T laboratories of Cambridge University. Here input images were in gray scale.

Another CNN based face recognition model is introduced in [4], here firstly detect the face and extract the feature from face and recognized. This model can be useful for surveillance, attendance taking and other small systems. They used AT&T dataset and gained 98.75% accuracy and gained 98.00% accuracy for their own real time customized dataset. In [5], a fusion of 3 different size CNN architectures have been introduced which are; CNN-S, CNN-M, CNN-L. With the size of CNN architectures size, number of convolution layers also increased here. Here they don't use dropout regularization layer as it didn't help their model to improved performance. LFW database is used to train the model. As different network captures the different regions information, so this fusion model boosts the performance of face recognition. Deep Neural Network has shown great result in the field of automatic visual facial feature extraction. In [6], a deep CNN based face recognition system with augmentation is introduced. To collect images, they created a separate website utilizing the AdaBoost algorithm and a skin color model, which students must login to, then select their faces and annotate with their ID. Face recognition has been fine-tuned using a VGG16 network pre-trained with a VGG-face dataset. This model acquired an accuracy of 86.03%. This kind a website login idea is also lengthy process though they claimed it as a time saving process than others. In [7], authors proposed a face recognition method using deep neural network instead of ConvNets. Their input was extracted facial features i.e., Haar cascade instead of raw pixel values for minimizing the complexity. They applied their proposed method on Yale Faces Dataset and their acquired average accuracy was 97.05%. It's a better approach for small dataset. Because of one additional step of extracting facial features, this method is not suitable for large datasets.

Hand gesture recognition have drawn an enormous attention of the researchers in recent years. Hand gesture recognition is like a blessing for deaf and other people who are unable to understand normal language. Expressive and meaningful body movements are known as gesture. Hand gesture recognition is now become a hot topic for researchers because of its use in various sectors like number recognition, virtual environment, graphic editor control, and television control. In [8], proposed a finger segmentation approach for hand gesture identification, in which the hand region is first recognized using background subtraction, and then the palm and fingers are divided to distinguish fingers. Hand motions are then classified using a simple rule classifier. Here a customized dataset of 1300 images for 13 gestures are used for training. Another dataset also used for comparison purpose. They applied this method for two different datasets and acquired average accuracy of 96.6%. Highly efficient for real time and no need of hand gloves. But the methods accuracy highly depends on the result of hand detection. Any kind of skin color moving objects other than hand can degrade the performance of the method.

A gesture recognition method using ConvNet is proposed in [9]. During pre-processing, they used morphological filters, contour creation, polygonal approximation, and segmentation to reduce noise and improve features in this method. They used several convolutional networks for training and testing, and their average success rate was 96 percent. It gives high success rate at a very low computational cost but this method can deal with only the gestures present in static images and it can't deal properly with occlusion problem. Here, in [10], a 3D CNN model is introduced which perform better than 2D CNN for HD-sEMG (High density surface EMG) based gesture recognition. 3D CNN is relatively simple than deep neural networks. They tested their architecture on CapgMyo DB-a and CSL-HDEMG datasets. For CapgMyo DB it shows 98% accuracy which is almost similar to 2D CNN but in case of CSL-HDEMG it shows far better performance than 2D CNN with 75% accuracy. Another 3D CNN based effective

method for dynamic hand gesture recognition have been introduced in [11]. A classifier based on a fused motion volume of normalized depth and image gradient values was proposed by the authors. It also makes use of spatiotemporal data augmentation to avoid overfitting. They used a challenging dataset which is VIVA dataset and their system acquired a good classification report which is 77.5%. In [12], a CNN based hand gesture recognition model is mentioned which is specially introduced for the deaf people's communication process. Here two CNN model used, one is for hand feature extraction and other one for upper body feature extraction. As a classifier ANN is used. Here dataset used for training is Chalearn Gesture dataset which consists of 20 different Italian gestures. An accuracy of 95.68% is gained from the model by using Nesterov's Accelerated Gradient descent (NAG). Deep CNN is now gaining a lot of attention due to easy feature extraction and many other benefits. In [13], introduced a Human Hand Gesture recognition system for identifying the American Sign Language (ASL).

Here, feature extraction is done by Deep Convolutional Neural Network (DCNN) and for classification process Multi-Class Support Vector Machine (MCSVM) is used. A customized dataset containing 26 signs from 3 different people were used for the model's evaluation and acquired an accuracy of 94.57%. In [14], a Deep Convolutional Neural Network (DCNN) based model for recognizing hand gesture is introduced which used in some mouse and keyboard operation for human computer interaction. Hand detection, hand tracking, and hand gesture recognition are discussed here. For feature extraction they use Deep Convolutional Neural Network (DCNN) and last layer of the network they use softmax activation function for classification task. Before entering the image into the CNN network, they applied several preprocessing with the initial images, such as Background Subtraction, Noise Processing, Skin Segmentation, Hand Detection using Haar Cascade Classifier, KCF tracker (for movement), Median-flow tracker (for zooming). And they got 98.44% accuracy for validation data set and 93.25%.

We have reviewed recent state of works on face and hand gesture recognition. Different shallow methods were used in face and hand recognition previously. But Deep CNN or CNN both are now on the center of attraction for the recognition type researches as it has easy feature extraction traits and it can learn without human supervision. Without these its cost and time efficient. It's giving good result for different problems even for complex one.

3. Explanation of the Proposed System

In this section, the details of the proposed method are explained. Our Person Identification system is based on two models i.e., Face and Hand Gesture Model.

3.1. Workflow Diagram of Face Recognition Model

The work flow diagram of our face model is presented in Fig. 1. From Fig. 1, after preprocessing of training data it will go to the Deep Learning model. The model will be learned based on training data. Then after preprocessing of testing data, it can predict the person using the learned model according to the labels.

3.1.1. Explanation of face recognition model

Face model is developed based on VGG16. VGG stands for Oxford University's Visual Geometric Group, and 16 refer to a 16-layer network.

VGG16 is a pretty large network with about 138 million parameters. VGG16 was proposed by Karen Simonyan and Andrew Zisserman in 2014 in the paper "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE_SCALE IMAGE RECOGNITION" [15]. It was one of the best performing architectures in ILSVRC challenge.

The network's input image has a dimension of (224, 224, 3). The initial two layers have 64 filter size convolution layers, followed by a max pool layer and two layers with 128 filter size convolution layers. It is followed by a stride (2, 2) max pooling layer. Then there are three more 256 filter convolution layers with the stride (3, 3) max pooling layer. Following that, there are two sets of three convolution layers, as well as a max pool layer with a filter size of 512. We acquire a feature map after stacking convolution and max-pooling, then we flatten it to transform it to a feature vector. It then went to the fully connected layer. Fig. 2 shows an illustration of the VGG16 model.

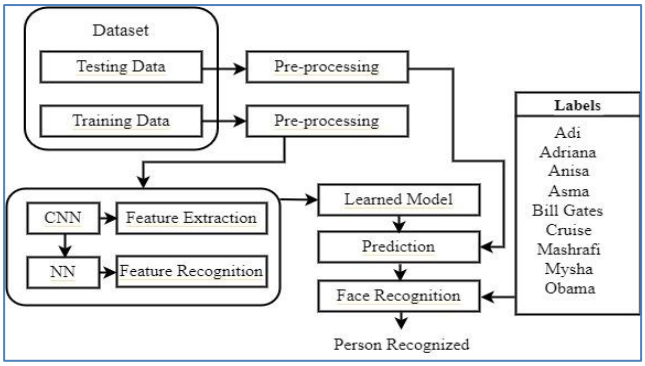


Fig. 1. Work Flow Diagram of Face Recognition Model.

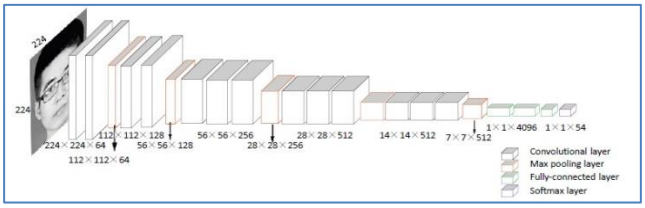


Fig. 2. An Illustration of VGG16 Model [6].

3.2. Workflow Diagram of Hand Gesture Recognition Model

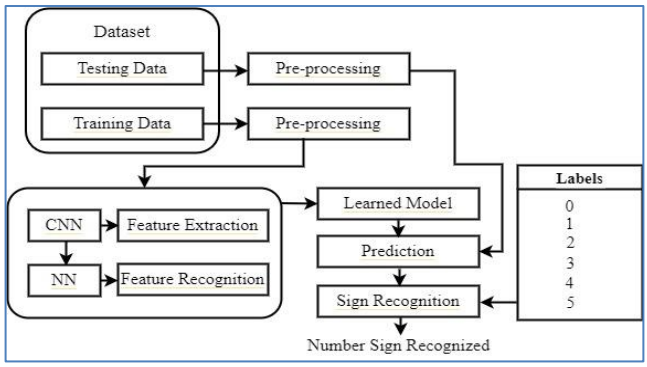


Fig. 3. Work Flow Diagram of Hand Gesture (Number Sign) Recognition model.

From the Fig. 3, we can see that after preprocessing of training data it will go to the deep learning model. The model will be learned based on training data. Then after preprocessing of testing data, it can predict the number using the learned model according to the labels.

3.2.1. Explanation of hand gesture recognition model

Convolutional Neural Network is used in the Hand Gesture Model (CNN). CNN (Convolutional Neural Network) is a deep learning algorithm that may be used to extract and classify features.

It is a multi-layer neural network. There are mainly two parts: one for feature extraction and another one for classification. In CNN architecture, there will be convolution layers, a fully connected layer, and a classification layer. Each convolution layers will be followed by an activation layer and a max pooling layer. An activation function in the activation layer assesses whether each neuron's input is meaningful for the model's prediction and whether it should be activated or not. The pooling layer minimizes the image's size. A perceptron layer is the Fully Connected (FC) layer. And the classification layer i.e., Softmax, which predicts the input image's class. For the classification task, there are two convolution layers, a pooling layer, and a fully - connected layer. Fig. 4 shows an illustration of a CNN model with two convolution layers.

The input size of the suggested CNN's model was set to 64x64x3. The kernel size for the first convolution layer was 3x3 and the number of filters was 32. Here, ReLU activation function was used as ReLU has shown the more efficiency than the alternative activation function [21]. By making the mapping function more flexible and non-linear, the ReLU helps prevent creating negative values, as well as avoiding saturation of the neurons [3]. The kernel size used for the max-pooling layers was 2x2. Afterwards, the number of filters used in the second convolution layer was 32, and the kernel size was the same as in the first convolution layer. After that, the features were flattened to make it into a feature vector. Furthermore, the features were moved to the fully connected layer which was mentioned as the dense layer. The dense implementation uses a 128-unit layer, which is followed by a final layer that computes the softmax probability for each of the six categories.

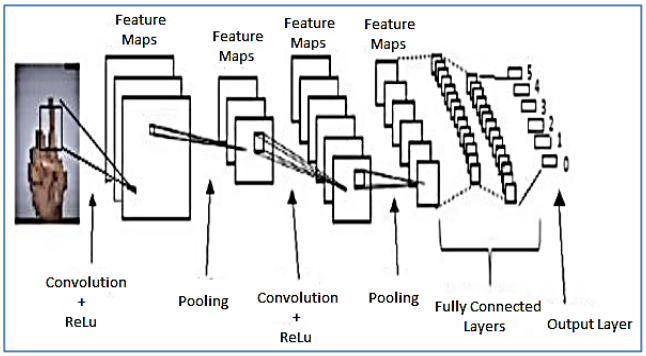


Fig. 4. An Illustration of CNN model with 2 Convolution Layer.

3.3. Basic Workflow Diagram of Proposed System

A Person Identification system is established based on Deep Learning. System firstly requests for a face image and cropped the face from the image. Then the cropped face image undergoes the face model and predicts the person. If the person's image is not matched with the Face Database the person will not be recognized. So, the procedure won't move forward and it will go back to the initial stage for requesting another face. But if the predicted person's image matched with the database, then it will request for a gesture image i.e., Number Sign, the number sign image will be cropped. This cropped gesture image will go through the Hand Gesture (Number Sign) Model and predict the sign. This predicted sign image goes to database and check if the gesture number sign is the one which is associated with the predicted person or not. If yes, then the person will be identified as "Authorized" and if no then the person will be identified as "Unauthorized". The block diagram of the proposed system is presented in Fig. 5.

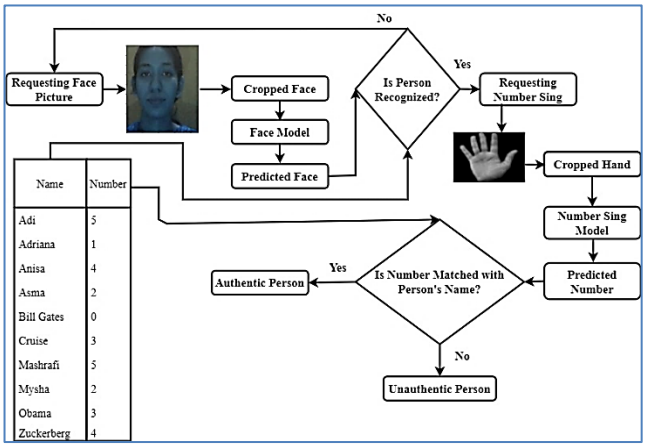


Fig. 5. Basic Workflow Diagram of the Proposed System

3.4. Classification Loss Function

The loss function evaluates a classification model's performance or translates decisions to their related costs. Loss function and cost function both are synonymous and use interchangeably. Here we used cross-entropy loss function. Depending on the context, cross-entropy and log loss are slightly different but both are considered as same and It assesses the performance of a model whose output is a probability value ranging from 0 to 1.

Here we used categorical cross-entropy for both of the model as our task is multi-class (where each sample can belong to only one class and the model must decide the which one is this class) classification. Actually, it's the quantifier of the difference between two probability distributions. By computing the following sum, it calculates the loss. For both models, we employed categorical cross-entropy as the loss function [2]. The loss function is presented in (1).

$$Loss = -\sum_{i=1}^{output\ size} y_i \,.\, log\hat{y}_i \qquad (1)$$

where, $\hat{y}_i$ is the i-th scalar value in the model output, y_i is the corresponding target value, output size is the number of scalar values in the model output. With categorical cross-entropy loss function, the Softmax activation function is recommended since it rescales the model output.

3.5. Optimization Algorithm

Optimizers are techniques or approaches that are primarily used to adjust the properties of a neural network, such as weights or learning rate, in order to lower the network's loss. Optimization Algorithm refers to a procedure that executed iteratively by comparing various solutions till a satisfactory solution is found. For Deep Learning model, choice of an optimization algorithm can be the reason of the models good or bad performance. One of the most basic and most used optimization algorithms is Gradient Descent. Gradient Descent mostly used in classification algorithm and linear regression. One of the most used and efficient variant of Gradient Descent algorithm is Adam optimizer. Here, we have used Adam optimizer for training both Face and Hand (Number Sign) recognition models.

Adam [16] stands for Adaptive Moment Estimation and it's the most modified technique of gradient descent. Adam mainly combines the best qualities of two algorithms named Adaptive Gradient Algorithm (AdaGrad) [17] and Root Mean Square Propagation (RMSProp) [18] and it provides easy configuration. It works with both first and second order momentum. Adam keeps

an exponentially decaying average of past gradient like AdaDelta. Adam calculates the momentum and learning rate for every parameter. Adam is too fast then other adaptive learning model and converges fast but it's also computationally costly like others.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \quad (2)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad (3)$$

where, $m_t, v_t =$ estimates of the first moment (the mean) and the second moment (the uncentered variance).
And,$\beta_1, \beta_2 =$decay rates of average of gradients $(\beta_1 = 0.9, \beta_2 = 0.999)$.
We used two Customized datasets for our face and number sign model training. Our face model based on VGG16 architecture and number sign model is a simple Neural Network with two convolutional layers. We used categorical cross-entropy as classification loss function as ours one is a multi-class problem. And used Adam optimizer as an optimization algorithm.

4. Experimental Results and Discussions

This section contains detailed experimental review of our proposed system. Here, we explain the implementation details of our face and hand model. And explain the confusion matrix, classification report and other experimental result in details.

4.1. Used Dataset

For Face dataset, we used some images that are downloaded from google and used them for making our face dataset. For number sign dataset, the "Sign Language for Numbers" dataset was download from Kaggle [19] having all the images. This file consists of 11 labeled folders for 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, unknown.

4.1.1. Face dataset

Our customized face dataset contains 10,000 images for 10 subjects. For each subject there are 1000 images. Here, we have collected real time 4000 images for 4 subjects. And we also have downloaded some images of other 6 subjects from internet. Some sample images are added in Fig.6 from each class of the dataset.

Fig. 6. Sample Images of Face Dataset.

4.1.2. Number sign dataset

Sign Language for Numbers dataset from Kaggle [19] contains 16500 images for 11 classes of American Sign Language images was collected which are 0-9 and unknown. We customized this dataset according to our need. We take 0-5 classes where each class contains 1500 images. So, our dataset contains total 9000 images for number signs. Some sample images of this dataset are shown in Fig. 7.

Fig. 7. Sample Images of Number Sign Dataset

4.2. Performance Analysis

We have trained our face model for different number of epochs for making the model more stable using our face dataset. We have trained our model for 5, 10, 15, 20 and 25 epochs and analyzed the accuracy and validation accuracy of the model. Here we used Adam optimizer as an optimization algorithm and categorical cross-entropy loss function as ours model is for solving classification type problem. Besides, we have trained our number sign model also for different number of epochs with our customized dataset. We have trained this number sign model for 5, 10, 15 and 20 epochs and analyzed the model's accuracy and validation accuracy. Here, also used Adam optimizer and categorical cross-entropy as a loss function.
Our dataset was divided into two parts: training and testing. The training set is used to develop and fit the model, whereas the testing set or validation set is used to determine the model's prediction accuracy. The accuracy and loss curves calculate and plot training loss, training accuracy, and validation loss, validation accuracy for our datasets on our models. The mistake in the training data is referred to as training loss, while the error after running the validation data set in the model is referred to as validation loss. Training accuracy refers to the correctness of training data, while testing accuracy refers to the accuracy of testing data.
We have also created a confusion matrix for both of our models. It's a table that shows how well a model performed on a test dataset. The model's prediction mistakes are shown in the confusion matrix. It has True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) values (FP). Classification report have also generated from confusion matrix. Here, it shows the models precision, recall, F1-score and accuracy, which are presented in (4) to (7) respectively [20]. These are described shortly here.
Precision: Precision refers to a classifier's ability to avoid mislabeling a negative occurrence as a positive one. The ratio of genuine positives to the sum of true positives and false positives is what it's called.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall: The ability of a classifier to detect all positive examples is defined as the ratio of true positives to the sum of true positives and false negatives for each class.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

F1-Score: The F1-Score is the weighted harmonic mean of recall and precision, with 1.0 being the highest and 0.0 being the worst.

$$F1 - Score = \frac{2*(Recall*Precision)}{Recall+Precision} \quad (6)$$

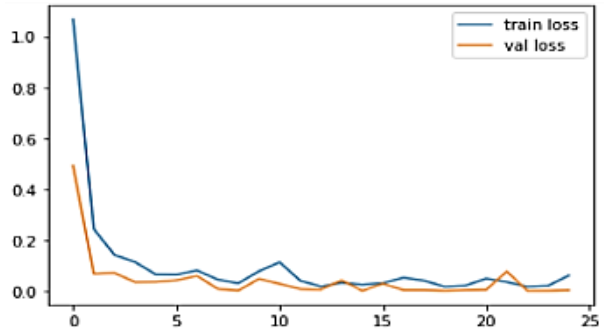
Support: The quantity of actual instances of the class in the supplied dataset is referred to as support, and it does not vary between models; rather, it diagnoses the assessment process.
Accuracy: The most basic performance measure of models is accuracy, which is defined as the ratio of properly predicted observations to total observations.

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (7)$$

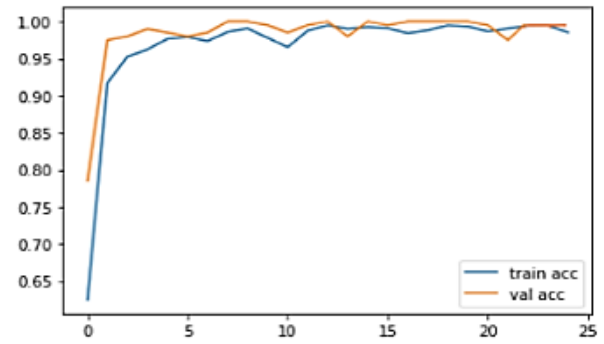
Following subsections are organized as, performance analysis of face model and hand gesture model with customized dataset.

4.2.1. Performance analysis of face recognition model

The training loss and the validation loss for face dataset by VGG16 is shown in Fig. 8(a). Here, training loss and validation loss are decreasing with the increasing number of epochs.



(a)



(b)

Fig. 8. For Face Dataset on Face Recognition Model: a) Training and Validation Loss Curve, and b) Training and Validation Accuracy Curve.

VGG16's training and validation accuracy for the face dataset are displayed in Fig.8 (b). When the number of epochs is increased, the training and validation accuracy improves. For 20 epoch we got 98.55% training accuracy and 97.50% validation accuracy. For 25 epoch we got 99.32% training accuracy and 98% validation accuracy. We train our model for 25 epoch and we got 98% validation accuracy for face dataset by VGG16.

The confusion matrix of the evaluated results is presented in Fig.9. We can see which face is predicted how well from the confusion matrix. Accuracy 98%, misclass 2% and predicted level for face recognition (Adi, Adriana, Anisa, Asma, Bill Gates, Cruise, Mashrafi, Mysha, Obama, Zuckerberg) which we can see from Confusion Matrix Fig.9. With respect to the confusion matrix, a classification report is given in Table 1. The complete support is 200. We have used 200 images for testing. Precision, Recall, F1 score and Support are calculated with respect to the confusion matrix.

Table 1. Classification Report of Face Recognition Model

Labels	**Precision**	**Recall**	**F1-score**	**Support**
Adi	1.00	1.00	1.00	20
Adriana	1.00	1.00	1.00	20
Anisa	1.00	1.00	1.00	20
Asma	1.00	1.00	1.00	20
Bill Gates	1.00	1.00	1.00	20
Cruise	0.95	1.00	0.98	20
Mashrafi	1.00	1.00	1.00	20
Mysha	1.00	1.00	1.00	20
Obama	1.00	0.80	0.89	20
Zuckerberg	0.87	1.00	0.93	20
Performance				
Accuracy			0.98	200
Macro Avg	0.98	0.98	0.98	200
Weighted Avg	0.98	0.98	0.98	200

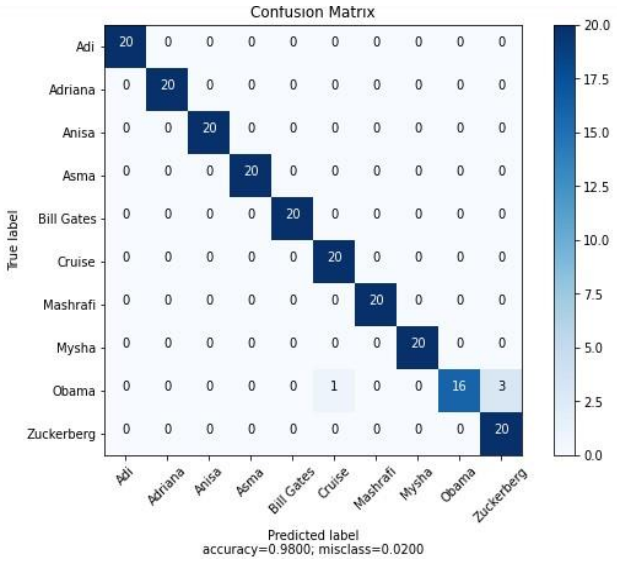
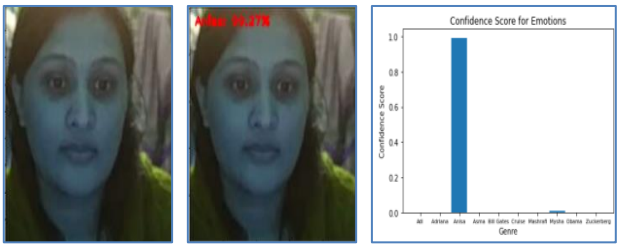
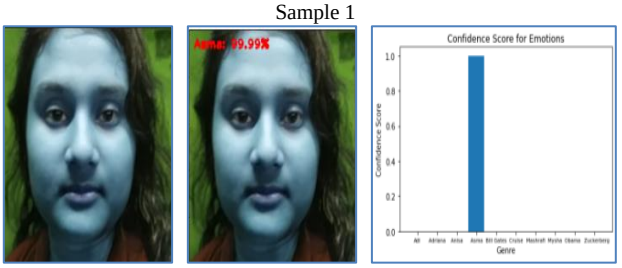


Fig. 9. Confusion Matrix of Face Recognition Model.

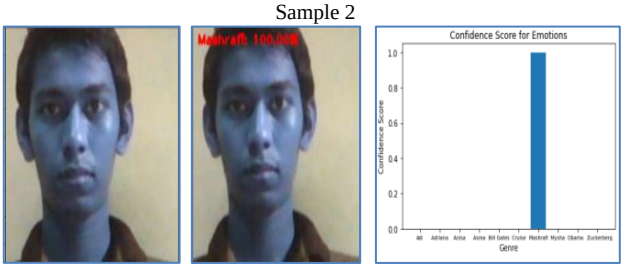
Here, individual persons were recognized by the face model very well. Fig.10 shows the confidence scores for each level.



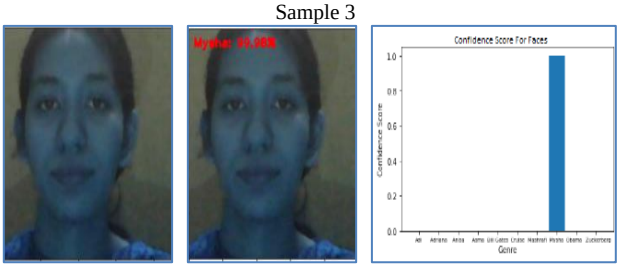
Sample 1



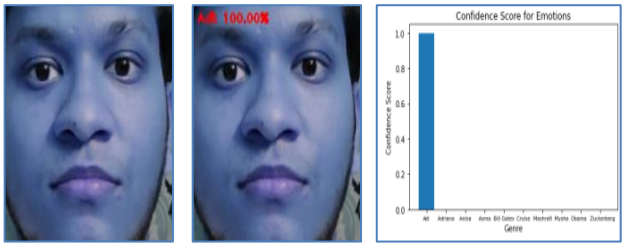
Sample 2



Sample 3



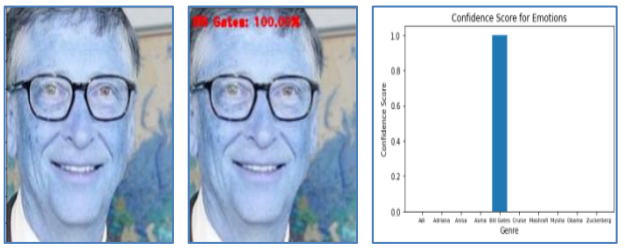
Sample 4



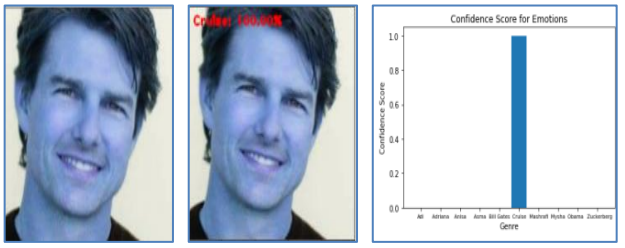
Sample 5

(a) (b) (c)

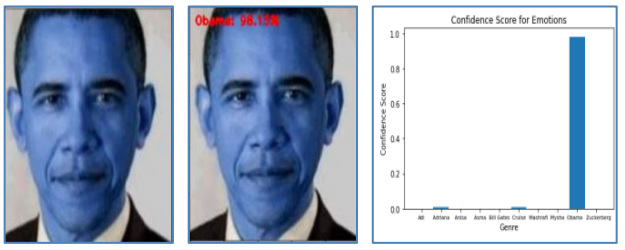
Fig. 10 (I). Experimental Result on Face Dataset for Face Recognition Model: a) Face Level, b) Face Recognition with Confidence Score, c) Confidence Scores in Bar chart.



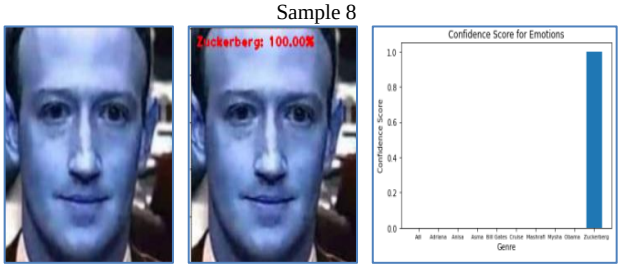
Sample 6



Sample 7



Sample 8



Sample 9

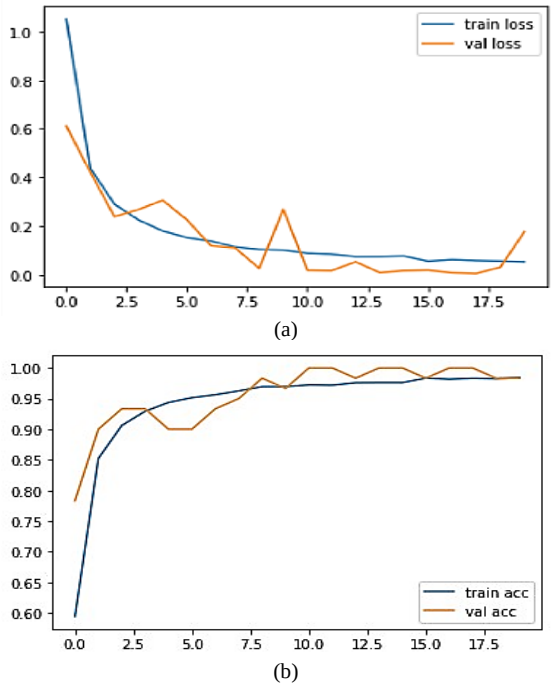
(a) (b) (c)

Fig. 10 (II). Experimental Result on Face Dataset for Face Recognition Model: a) Face Level, b) Face Recognition with Confidence Score, c) Confidence Scores in Bar chart.

In Fig. 10(I) & Fig. 10(II), sample 1 and sample 2 shows 99.27% & 99.99% confidence score for Anisa and Asma. Sample 3 shows 100% for Mashrafi and 99.98% for Mysha in sample 4. 100% score shows for Adi in sample 5. 100 % confidence scores for Bill gates and Cruise in Sample 6 and 7. Sample 8 and 9 shows 99.15%, 100 % score respectively for Obama and Zuckerbarg.

4.2.2. Performance analysis of hand gesture i.e., number Sign recognition model

The training loss and the validation loss for our number dataset of proposed method is shown in Fig. 11(a). Here, training loss and validation loss are decreasing with the increasing number of epochs.



(a)

(b)

Fig. 11. For Number Sign Dataset on Hand Gesture Recognition Model: a) Training and Validation Loss Curve b) Training and Validation Accuracy Curve.

Fig. 11(b) depicts the training and validation accuracy for our number dataset using the proposed strategy. When the number of epochs is increased, the training and validation accuracy improves. For 5 epochs, we got 94.37% training accuracy and 90% validation accuracy. For 10 epochs, we got 96.96% training accuracy and 96.67% validation accuracy. For 20 epoch we got 98.42% training accuracy and 98.33% validation accuracy. So, we got 98.33% validation accuracy of number sign based proposed method for our number sign dataset. The evaluated result of confusion matrix is presented in Fig. 12. We can see which number predicted how well from the Confusion Matrix. Accuracy 98.33%, misclass 1.67% and predicted level for number sign (0, 1, 2, 3, 4, 5) which we can see from confusion matrix Fig.12.

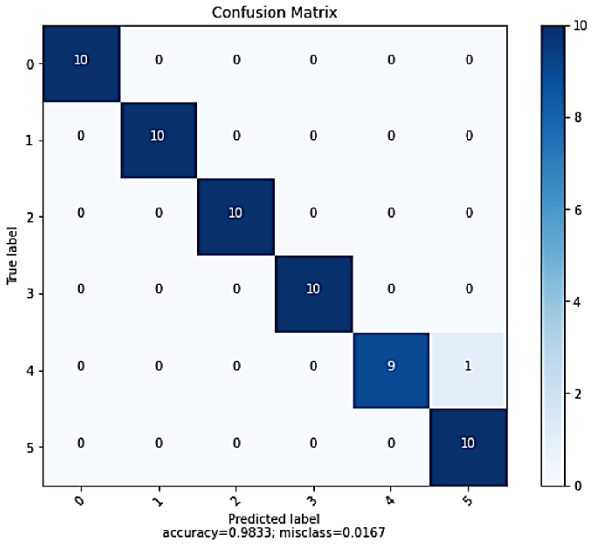


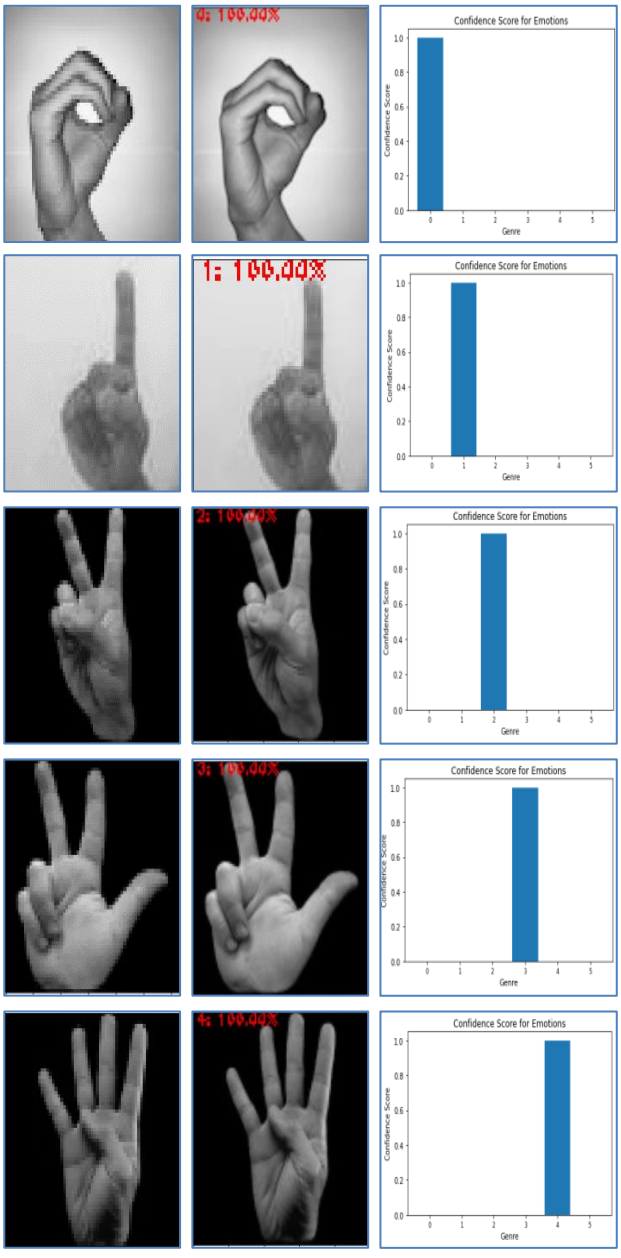
Fig. 12. Confusion Matrix of Hand Gesture Recognition Model.

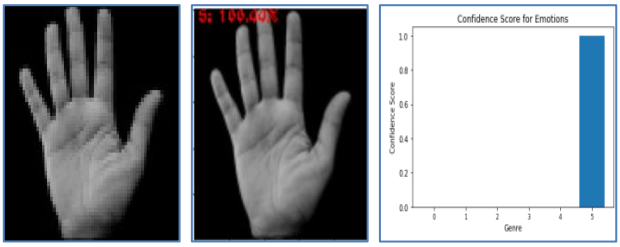
With respect to the confusion matrix, a classification report is given in Table 2. The complete support is 60. We have used 60 images for testing. Precision, Recall, F1 score and Support are calculated with respect to the confusion matrix.

Table 2. Classification report of hand gesture i.e., number sign recognition model

Labels	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	10
1	1.00	1.00	1.00	10
2	1.00	1.00	1.00	10
3	1.00	1.00	1.00	10
4	1.00	0.90	0.95	10
5	0.91	1.00	0.95	10
Performance				
Accuracy			0.98	60
Macro Avg	0.98	0.98	0.98	60
Weighted Avg	0.98	0.98	0.98	60

Here, individual gestures were recognized by our hand model very well. Fig.13 shows the confidence scores for each level.





(a) (b) (c)

Fig. 13. Experimental Result on Number Sign Dataset for Hand Gesture (Number Sign) Recognition Model: a) Number sign Level, b) Number Sign Recognition with Confidence Score, c) Confidence Scores in Bar chart

From Fig.13, we can see that sample 1 and sample 2 shows 100% confidence score for number sign 0 and 1. 100% score also shows for number sign 2 in sample 3 and shows 99.67% for sign 3 in sample 4. Sample 5 and 6 also shows 100 % confidence scores for number sign 4 and number sign 5.

4.3. Overall Experimental Result of Our System

We have set some numbers signs as Hand Gesture with every individual of the dataset according to the following chart.

Table 3. Chart of Associated Number Sign with Every Individual

Person Name	Number
Adi	5
Adriana	1
Anisa	4
Asma	2
Bill Gates	0
Cruise	3
Mashrafi	5
Mysha	2
Obama	3
Zuckerberg	4

For performing the task of person identification, we have used “google colab” instead of any specific application. Where, we have used the pretrained Face and Hand Gesture model to identify the person as described in section 3.3 along with Fig.5. Some sample performance evaluation processes of person identification are shown in Fig. 14.

(a) (b) (c)

Fig. 14. Performance evaluation process of person identification: a) person recognition, b) number sign recognition, and c) final result of person authentication.

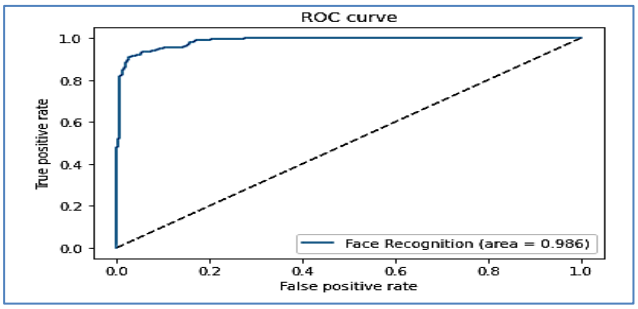
4.4. Discussions

In our proposed system, both face model and number sign model give better result using simple model architecture and datasets of almost 10,000 images (per dataset). And both of the models give good confidence scores also.

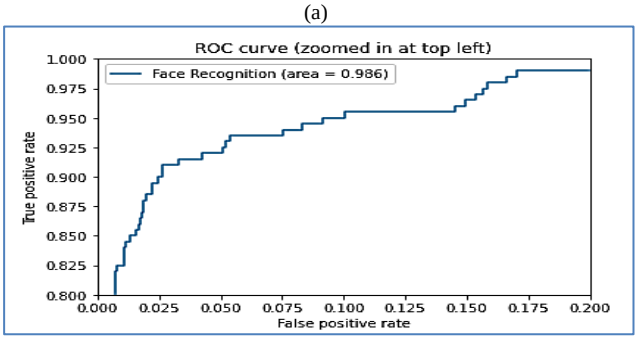
From Literature review of different face recognition model, we can see that, in [1], they used LBP and eigen face technique for performance enhancement. A CNN based model proposed in [4], which acquired 98.00% accuracy on real dataset but added an additional step for feature extraction. In [5], authors proposed a fusion model of 3 different sizes CNN model. And A Deep Neural Net model proposed in [7], which uses Yale Face database which is a small database and got 97.05% accuracy. So, in comparison with these complex techniques and deep models we used a VGG16, a CNN based model that gives 98.00% validation accuracy for our customized dataset of 10,000 images (for 10 subjects).

The performance of a classification task is measured using Receiver Operator Characteristics Curve (ROC). The true positive rate (TPR) is plotted against the false positive rate (FPR) to create a ROC curve (FPR). ROC curves are commonly used in binary classification to examine a classifier's output. Binarization of the output is required to expand the ROC curve and ROC area to multi-label classification [22].

For proposed Face Recognition model ROC curve is shown in Fig. 15(a) and ROC curve zoomed in at top left is shown in Fig. 15(b). And, from reviewing different research on hand gesture recognition model we can see that, in [8], authors used finger segmentation method where various complex steps are included for finger region detection and recognition, here they got 96.6% accuracy. A different convnet model is proposed in [9], that acquired 96% accuracy but applied different techniques for noise reduction. In [10], a 3D CNN model is introduced for HD-sEMG based gesture recognition and shows 98.00% accuracy on CapgMyo DB. Another 3D-CNN model acquired accuracy of 77.5% on VIVA dataset in [11]. In [12], a CNN based model got accuracy of 96.5% but used two different model used for different regions feature extraction. And another model is proposed in [13], which uses DCNN for feature extraction and MCSVM for classification and got 94.75% accuracy for American Sign Language (ASL). In this thesis work we used a simple CNN model with 2 convolution layer which shows 98.33% accuracy for our customized Number Sign Dataset without any complex techniques or methods. For proposed Hand Gesture (Number Sign) model ROC curve is shown in Fig. 16(a) and ROC curve zoomed in at top left is shown in Fig. 16(b).



(a)

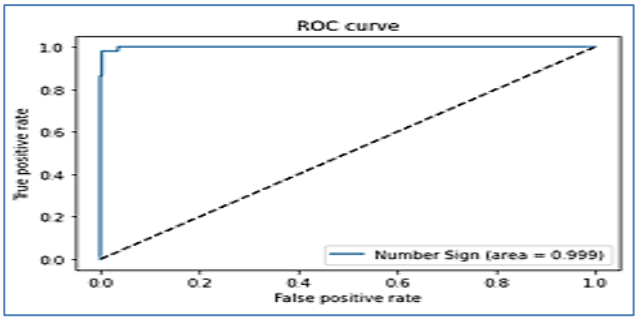


(b)

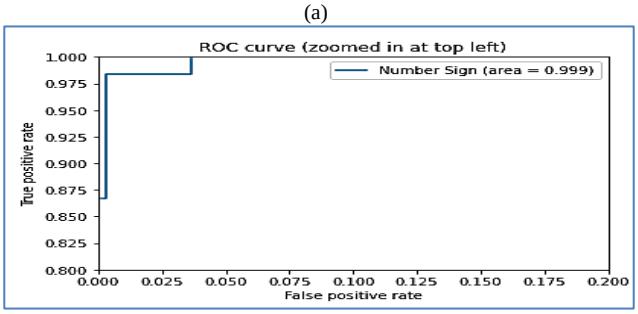
Fig. 15. For proposed Face Recognition Model: a) ROC curve, and b) ROC curve zoomed in at top left

Table 4. Accuracy comparison table for face recognition model

Model Name	Dataset	Accuracy
CNN [3]	ORL Dataset	98.7%
CNN [4]	AT&T Dataset	98.75%
DCNN [6]	Customized Dataset	86.03%
DNN [7]	YALE Face Dataset	97.05%
Our Model (VGG16)	**Customized Dataset**	**98.00%**



(a)



(b)

Fig. 16. For proposed Hand Gesture (Number Sign) Recognition Model a) ROC curve b) ROC curve zoomed in at top left.

Table 5. Accuracy Comparison table for hand gesture i.e., number sign recognition model

Model Name	Dataset	Accuracy
Finger segmentation [8]	Customized Dataset	96.6%
ConvNet [9]	Customized Dataset	96%
3D-CNN [10]	Capg-Myo DB CSL-HDEMG DB	98% 75%
3D-CNN [11]	VIVA Dataset	77.5%
CNN [12]	Chalearn Dataset	95.68%
DCNN [13]	ASL	94.75%
Our Model (CNN-2 Convolution layers)	**Customized Dataset**	**98.33%**

So, on overall performance analysis and experiment result of our both models we can say that our face model and hand gesture (number sign) model acquired a good accuracy rate with and customized simple model dataset in comparison of reviewed model of different research papers.

5. Conclusions

Security means freedom or protection from any kind of harm caused by any unwanted person. Person Identification for security system is now gaining a lot of attention because everyone is very much conscious about their security. Till now person identification was based on face recognition, biometric based finger recognition, iris recognition etc. So, we proposed a new security system which will identify person based on both face feature and number sign gesture i.e., hand simultaneously. Our proposed Person Identification system contains two simple models. One is VGG16 based Face Model and a CNN (2 convolution layers) based Hand Gesture (Number Sign) Model. Both models give 98.00% and 98.33% accuracy respectively. These simple models of proposed system show better accuracy than many complex models. In our system, two recognition models have been merged for making a security system stronger and two average size customized datasets have been used for two simple network-based model. It acquired a satisfactory result. We plan to develop a new dataset with more classes for further improvement.

References

[1] B. Kranthikiran, and P. Pulicherla, "Face Detection and Recognition for use in Campus Surveillance," International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 3, pp. 2908–2913, Jan. 2020, doi: 10.35940/ijitee.b7682.019320.

[2] M. Coşkun, A. Uçar, Ö. Yildirim, and Y. Demir, "A FACE RECOGNITION ALGORITHM BASED ON CONVOLUTIONAL NEURAL NETWORK," REVISTA ARGENTINA DE CLINICA PSICOLOGICA, 2020, doi: 10.24205/03276716.2020.339.

[3] Y. Said, M. Barr, and H. E. Ahmed, "Design of a Face Recognition System based on Convolutional Neural Network (CNN)," Engineering, Technology & Applied Science Research, vol. 10, no. 3, pp. 5608–5612, Jun. 2020, doi: 10.48084/etasr.3490.

[4] K. B. Pranav, and J. Manikandan, "Design and Evaluation of a Real-Time Face Recognition System using Convolutional Neural Networks," Procedia Computer Science, vol. 171, pp. 1651–1659, Jan. 2020, doi: 10.1016/j.procs.2020.04.177.

[5] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, & T. Hospedales, "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition," In Proceedings of the IEEE international conference on computer vision workshops, pp. 142-150, IEEE, 2015.

[6] Z. Pei, H. Xu, Y. Zhang, M. Guo, and Y.-H. Yang, "Face Recognition via Deep Learning Using Data Augmentation Based on Orthogonal Experiments," Electronics, vol. 8, no. 10, p. 1088, Sep. 2019, doi: 10.3390/electronics8101088.

[7] P. Gupta, N. Saxena, M. Sharma, and J. Tripathi, "Deep learning model for group face recognition based on Convolution Neural Network," Journal of Xidian University, vol. 14, no. 5, May 2020, doi: 10.37896/jxu14.5/415.

[8] Z. Chen, J.-T. Kim, J. Liang, J. Zhang, and Y.-B. Yuan, "Real-Time Hand Gesture Recognition Using Finger Segmentation," The Scientific World Journal, vol. 2014, p. e267872, Jun. 2014, doi: 10.1155/2014/267872.

[9] R. F. Pinto, C. D. B. Borges, A. M. A. Almeida, and I. C. Paula, "Static Hand Gesture Recognition Based on Convolutional Neural Networks," Journal of Electrical and Computer Engineering, vol. 2019, pp. 1–12, Oct. 2019, doi: 10.1155/2019/4167890.

[10] J. Chen, S. Bi, G. Zhang, and G. Cao, "High-Density Surface EMG-Based Gesture Recognition Using a 3D Convolutional Neural Network," Sensors, vol. 20, no. 4, p. 1201, Feb. 2020, doi: 10.3390/s20041201.

[11] P. Molchanov, S. Gupta, K. Kim, & J. Kautz, "Hand gesture recognition with 3D convolutional neural network," In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 1-7, IEEE, 2015.

[12] L. Pigou, S, Dieleman, P. J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," In European Conference on Computer Vision, pp. 572-578, Springer, Chamm, 2014, September.

[13] M.R. Islam, U.K. Mitu, R.A. Bhuiyan, and J. Shin, "Hand gesture feature extraction using deep convolutional neural network for recognizing American Sign Language," In 2018 4th International Conference on Frontiers of Signal Processing (ICFSP), pp. 115-119, IEEE, 2018, September.

[14] M. M. Islam, M. R. Islam, and M. S. Islam, "An Efficient Human Computer Interaction through Hand Gesture Using Deep Convolutional Neural Network," SN Computer Science, vol. 1, no. 4, Jun. 2020, doi: 10.1007/s42979-020-00223-x.

[15] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[16] D. P. Kingma, & J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[17] J. Duchi, E. Hazan, & Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," Journal of Machine Learning Research, Vol.12, 2121–2159, 2011.

[18] S. Ruder, "An overview of gradient descent optimization algorithms," *ArXiv, abs/1609.04747*, 2016.

[19] https://www.kaggle.com/muhammadkhalid/sign-language-for-numbers (21 February, 2021).

[20] I. Cinar, and M. Koklu, "Classification of rice varieties using artificial intelligence methods," International Journal of Intelligent Systems and Applications in Engineering, 7(3), pp.188-194, 2019.

[21] A. Krizhevsky, I. Sutskever, & G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, *25*, 2012.

[22] Y. S. Taspinar, I. Cinar, & M. Koklu, "Prediction of Computer Type Using Benchmark Scores of Hardware Units," *Selcuk University Journal of Engineering Sciences*, *20*(1), 11-17, 2021.