Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

Facial Expression Recognition Through Transfer Learning: Integration of VGG16, ResNet, and AlexNet with a Multiclass Classifier

Balaiah Paulchamy[1*], Abid Yahya[2], Natarajan Chinnasamy[3], Kalpana Kasilingam[1]

[1] Department of Electronics and Communication Engineering, Hindusthan Institute of Technology, 641032 Coimbatore, India
[2] Department of Electrical Computer and Telecommunication, Botswana University of Science and Technology, PO BOX 016 Palapye, Botswana
[3] Department of Mechanical Engineering, Hindusthan Institute of Technology, 641032 Coimbatore, India

[*] Correspondence: Balaiah Paulchamy (drpaulchamy@hit.edu.in)

Received: 12-24-2024 **Revised:** 02-10-2025 **Accepted:** 02-16-2025

Citation: B. Paulchamy, A. Yahya, N. Chinnasamy, and K. Kasilingam, "Facial expression recognition through transfer learning: Integration of VGG16, ResNet, and AlexNet with a multiclass classifier," *Acadlore Trans. Mach. Learn.*, vol. 4, no. 1, pp. 25–39, 2025. https://doi.org/10.56578/ataiml040103.

Abstract: This study investigates the recognition of seven primary human emotions—contempt, anger, disgust, surprise, fear, happiness, and sadness—based on facial expressions. A transfer learning approach was employed, utilizing three pre-trained convolutional neural network (CNN) architectures: AlexNet, VGG16, and ResNet50. The system was structured to perform facial expression recognition (FER) by incorporating three key stages: face detection, feature extraction, and emotion classification using a multiclass classifier. The proposed methodology was designed to enhance pattern recognition accuracy through a carefully structured training pipeline. Furthermore, the performance of the transfer learning models was compared using a multiclass support vector machine (SVM) classifier, and extensive testing was planned on large-scale datasets to further evaluate detection accuracy. This study addresses the challenge of spontaneous FER, a critical research area in human-computer interaction, security, and healthcare. A key contribution of this study is the development of an efficient feature extraction method, which facilitates FER with minimal reliance on extensive datasets. The proposed system demonstrates notable improvements in recognition accuracy compared to traditional approaches, significantly reducing misclassification rates. It is also shown to require less computational time and resources, thereby enhancing its scalability and applicability to real-world scenarios. The approach outperforms conventional techniques, including SVMs with handcrafted features, by leveraging the robust feature extraction capabilities of transfer learning. This framework offers a scalable and reliable solution for FER tasks, with potential applications in healthcare, security, and human-computer interaction. Additionally, the system's ability to function effectively in the absence of a caregiver provides significant assistance to individuals with disabilities in expressing their emotional needs. This research contributes to the growing body of work on facial emotion recognition and paves the way for future advancements in artificial intelligence-driven emotion detection systems.

Keywords: Convolutional neural network; Feature extraction; Facial expression recognition; Support vector machine classifier; Transfer learning

1 Introduction

There is an active study in facial analytics, and the key features of facial expressions, such as poses, expressions, gender, age, identity, etc., are extracted. Rule enforcement, active device authentication, face biometrics for payments, self-driving cars, and other uses are only a few of its many applications [1]. Reactions are the body's emotional responses to something or someone through modifying its behaviour and physiological functions. Emotion recognition enables the identification of basic human emotions, including fear, repulsion, sadness, happiness, and amazement, based on certain input information [2]. The movable parts of the face ultimately create expressions on the face. Expressions vary from person to person. Applications for emotion recognition can be found in human surveillance, advertising, e-learning, medical care, entertainment, and protection.

Machine learning (ML) is the study of algorithms that may automatically improve efficiency, data prediction, data learning, and accuracy via practice and the use of data. ML is beneficial for accurately detecting and recognizing features with the help of techniques. ML has a wide range of applications in healthcare, vision in computers, speech recognition, and image recognition. Models for emotion recognition that can be effectively trained to recognize human emotions can be created using ML algorithms [3].

Deep learning (DL), a branch of ML, fully uses artificial neural networks to translate particular technical concepts from the human brain to a computer. Specifically, categories can be learned sequentially because of the hidden layer concept of the DL technique. When all of the neurons or nodes in the network are joined, an accurate visual representation of the network as an ensemble is produced. It is helpful that DL does not require human feature extraction when using an algorithm to extract essential properties [4]. To determine human emotions in images, the CNN algorithm was used. It is responsible for feature extraction and image classification based on facial features. Overall, the suggested method, which integrates transfer learning, Deep convolutional neural networks (DCNNs), and SVM classifiers, successfully addressed the drawbacks of previous FER research, including poor generalization, overfitting, limited feature extraction, and high processing needs. As a result, the approach becomes more precise, effective, and flexible for practical uses, paving the way for further developments in FER technology.

2 Related Works

Various applications, such as social robots, video games, and human-machine interaction in security surveillance, have exploited FER [5]. Medical science employs FER to track pain, depression, and anxiety and treat dementia, whereas behavioural research uses it to gather social information like age, gender, and origin. Most facial expressions are easily visible to humans, but accurate computer expression identification is still challenging. Appropriate pre-processing, feature extraction, and image classification are the three key challenges in FER, especially when dealing with changeable input data settings, the head pose, cluttered environments, irregular illumination, and different causes of variation in the face. Despite DL's ability to absorb features, FER still has difficulties. First, to avoid overfitting, deep neural networks require a large amount of training data [6].

The commonly recognized deep neural network framework, which provides the greatest achievable rate of object recognition task accuracy, cannot be trained due to the lack of facial expression databases. Due to numerous personal characteristics like background, gender, age, ethnicity, and level of expression, there is also a considerable inter-subject variation [7]. In addition to subject identity bias, other factors that affect how a face is perceived include lighting effects, occlusions, and differences in position variability. It helps in improving the deep network requirement to train active expression precise representational symbols and overcome the variation in the huge intra-class since none of them is longitudinally associated with facial expressions.

A feature extraction, face identification, and classification method that utilizes transfer learning feature extraction methods, such as VGG19, AlexNet, and ResNet architecture, was proposed to enhance FER performance and address the above difficulties. DL techniques, particularly CNN frameworks, which are multistage, anatomically motivated ones that autonomously train networks of continuous features, generated notable results when employed for feature extraction and classification [8]. The ConvNets have a high-rank representative feature and multistage input picture processing for hierarchical extraction. Jung et al. [9] created an efficient method that uses ConvNets to recognize facial expressions. By utilizing CNN models, facial expressions were classified into discrete categories, such as happiness, anger, surprise, sadness, and disgust, based on input image data.

Previous techniques would immediately transfer dark colour to the reference image, ignoring the facial emotion in the image, if the background of the targeted image is dark. In contrast, the target image has a happy expression on its face. Liu et al. [10] suggested an additional paradigm for emotional colour transfer that considers facial expressions. The face emotion label of a target image comprising elements of a facial expression was originally predicted using the emotion classification network. Then, pre-trained emotional colour transfer models were linked with facial emotion labels. The target image's color was then transferred to the source image using the matching emotion description. To regulate the FaceChannel, Barros et al. [11] presented a lightweight neural network with significantly fewer parameters than conventional deep neural networks. Automatic FER was based on deep NN, which is efficient but quite expensive to train. Thus, applying an inhibitory layer shaped the learning of facial features in the network's final layer, enhancing performance while lowering the number of parameters that can be trained. Akhand et al. [12] put forth the ensemble of numerous CNN-based FER approaches. In this technique, a face detector first extracted the face region from the pre-processed image. Second, after five key points were found for each image, two eye center points aligned the facial images. Third, three CNNs were trained for the entire face, eye, and mouth areas, one at a time, after the face picture was cropped into localized eye and mouth portions. The classification was then created using an ensemble of three CNNs' outputs. With inadequate information, transfer learning is an effective method for addressing categorization problems. For person recognition based on ear pictures, Almisreb et al. [13] used AlexNet CNN transfer learning. To identify ten classes rather than 1,000 classes, a fine-tuned AlexNet CNN was used, and the final completely connected layer was substituted with another fully connected (FC) layer. A Rectified Linear Unit (ReLU) layer was included to enhance the network's capacity for non-linear problem-solving. A database of people is necessary to train a neural network for face recognition using DL [14]. The trained network can then recognize objects. Pre-trained CNN has been used in the framework of the facial recognition process utilizing the transfer learning approach AlexNet.

Hernandez-Ortega et al. [15] suggested a DL-based quality assessment method for facial recognition. The technique used FaceQnet and CNN to forecast whether a given input image would be suitable for facial recognition. The VGGFace2 database was used for FaceQnet training. FaceNet was used to collect the baseline quality labels to produce comparison scores. The application of a refined ResNet-based CNN enabled it to return a numerical quality measure for each input image. Facial masks create a difficult issue for facial recognition programmes, which are frequently employed for security verification purposes, as these programmes are normally trained with human faces without masks. Still, now that the COVID-19 pandemic has started, they must detect faces with masks. Therefore, this problem can be solved by creating a DL-based model that accurately recognizes people wearing face masks. The network that excels at identifying masked faces can be trained using a ResNet-50-based architecture [16]. Prakash et al. [17] suggested using CNN with the transfer learning algorithm VGG16 for automated facial recognition. The images from the face database were trained using the CNN with weights learned from the pre-trained model VGG16 on the massive ImageNet database [18]. The FC layer and SoftMax activation were fed the retrieved features as input for classification. Table 1 summarises various research findings in FER. Each row represents a study. In contrast to other methods that prioritize accuracy, this work also highlights computational efficiency, demonstrating how transfer learning lowers the need for training time and resources. To lessen the impact of smaller and unbalanced FER datasets—a problem that has restricted the generalizability of previous models—it also integrates sophisticated data augmentation and balancing procedures. In addition to improving recognition accuracy, this research expands the scope of FER's applicability in real-world contexts like healthcare, security, and human-computer interaction by showcasing practical applications and exhibiting superior performance across common benchmarks [19, 20].

Table 1. Overview of research findings in FER

References	Methodology	Research Findings
[5]	Various applications, such as social robots, video games, and humanmachine interaction in security surveillance, have exploited FER.	FER was used in medical science to track pain, depression, and anxiety and treat dementia. It was also used in behavioral research to gather social information such as age, gender, and origin.
[6]	Lack of facial expression databases and inter-subject variation	The lack of facial expression databases prevents training deep neural networks for FER. There is considerable variation among individuals due to personal characteristics and other factors.
[8]	Transfer learning using VGG19, AlexNet, and ResNet architecture	DL techniques improve FER performance, particularly CNN models like VGG19, AlexNet, and ResNet.
[9]	CNN models	CNN models were used to predict facial expression labels.
[10]	Emotional color transfer	An emotional color transfer method was proposed that considers facial expressions.
[11]	Light-weight neural network with inhibitory layer	A lightweight neural network with an inhibitory layer improves performance in FER.
[12]	Ensemble of CNN-based FER approaches	An ensemble of CNNs for different facial areas improves classification in FER.
[13]	Transfer learning using AlexNet CNN for personal recognition	Transfer learning with an AlexNet CNN enhances person recognition based on ear pictures.
[14]	DL-based face recognition using pre-trained CNN	DL-based face recognition using a pretrained CNN achieves high accuracy.
[15]	DL-based quality assessment for facial recognition	A refined CNN-based method provides a numerical quality measure for facial recognition.
[16]	DL-based model for recognizing people wearing face masks	A DL-based model accurately recognizes people wearing face masks.
[17]	Automated facial recognition using CNN with the transfer learning algorithm	CNN with the VGG16 transfer learning algorithm achieves automated facial recognition.

3 Methodology

A DL model for FER using CNN and multiclass classification was developed in this study. Three renowned transfer learning models, namely, VGG16, ResNet, and AlexNet, were employed for feature extraction. A multiclass SVM classifier was used to augment the DL model to identify emotions precisely using facial data. The dataset, annotated and categorized into seven emotional states—contempt, anger, disgust, surprise, fear, happiness, and sadness—provides the raw material for the model.

Figure 1 provides a visual overview of the proposed methodology, illustrating the workflow. Each block represents a step in the process, showing the trajectory from data input to final emotional identification.

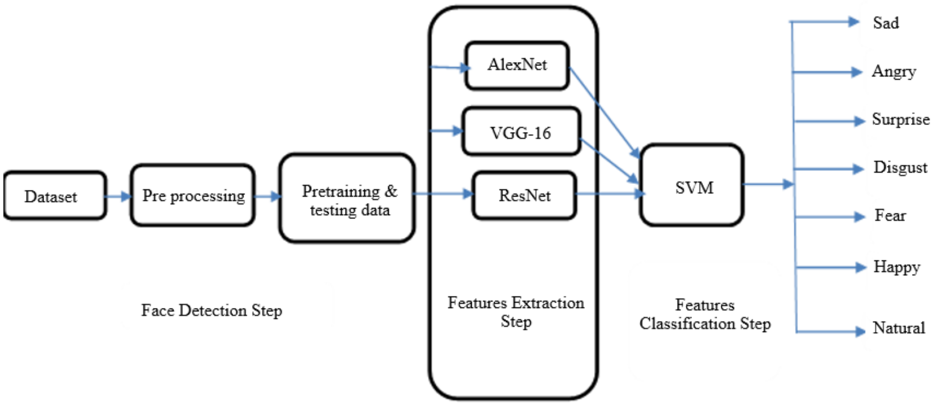


Figure 1. Block diagram of the proposed methodology

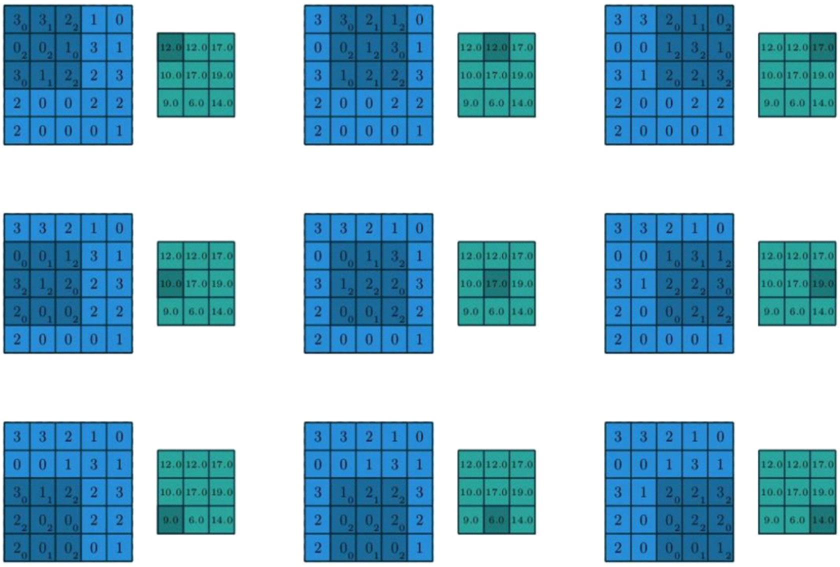


Figure 2. Convolution operation

3.1 Components of CNN

The CNN model is based on four major components or layers: convolutional layer, pooling layer, dropout layer, and FC layer.

(a) Convolutional layer

The convolutional layer is the primary layer, extracting distinctive attributes from the input images. By performing mathematical convolutions with a particular sized M×M filter on the input image, this layer identifies the dot product among the filter and different regions of the image. The processing results in a feature map containing information about the image's corners and edges, which subsequent layers can utilize to discern other image characteristics (Figure 2).

(b) Pooling layer

Following the convolutional layer, the pooling layer helps reduce the convolved feature map, saving computational resources. It minimizes connections between layers while operating autonomously on each feature map. Multiple pooling methods exist, including max pooling, average pooling, min pooling, and sum pooling. The pooling layer generally links the convolutional layer and the FC layer (Figure 3).

(c) Dropout layer

Dropout is a technique employed for regularizing ML algorithms. The dropout layer's purpose is randomly deactivating input units after each training cycle. An ensemble of 2n smaller networks trained like an ensemble of networks with dropout is realized, sharing weights to maintain a constant overall number of parameters. The effect of aggregating the predictions of these smaller networks can be approximately determined at test time using a single network with dropout and fewer weights (Figure 4).

(d) FC layer

The FC layer consists of neurons, weights, and biases and connects the neurons between two layers. Often placed before the final layers of a CNN architecture, this layer performs extensive computations and data analysis. The input images from the previous layers are flattened and passed onto the FC layer, which then performs standard mathematical computations on the flattened vector via some additional FC layers, thus initiating the classification process.

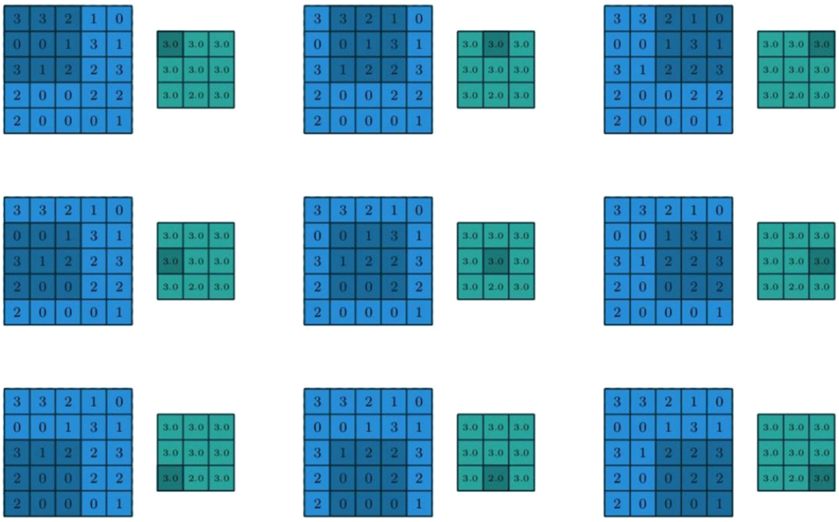


Figure 3. Max pooling operation

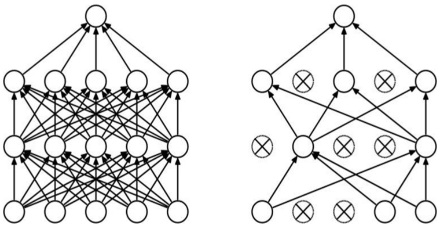


Figure 4. Densely connected neural network - dropout layer

3.2 Multiclass Classification

Multiclass classification refers to classifying instances into one of three or more classes. Given real-world data's unpredictable and intricate nature, developing automatic computer-based classification systems for data classification is increasingly relevant. For this case, a multiclass SVM classifier was leveraged to identify features extracted from seven emotion classes—contempt, anger, disgust, surprise, fear, happiness, and sadness.

3.3 Algorithm 1: FER

The proposed methodology employs Algorithm 1 for FER, which integrates CNN and a multiclass SVM to perform the task. This algorithm helps precisely identify various emotional states based on facial expressions, significantly contributing to emotion detection and recognition.

Algorithm 1 FER using CNN and a multiclass SVM

```
(a) Input
$D_{train} = \{I_i, E_i\}_{i=1}^{N}$: Pre-processed training dataset of facial expressions, where $I_i$ is the $i$-th image and $E_i$ is the corresponding emotion label.
$D_{test} = \{I_j, E_j\}_{j=1}^{M}$: Pre-processed testing dataset of facial expressions, where $I_j$ is the $j$-th image and $E_j$ is the corresponding emotion label.
(b) Output
$A$: Accuracy of the facial expression recognition model
(c) Procedure
Step 1: Initialize the CNN model:
  Define the CNN model $M$ with layer $L = \{L_1, L_2, \ldots, L_k\}$, where each layer $L_i$ performs a specific operation (convolution, pooling, dropout, fully connected, or output).
Step 2: Choose transfer learning models for feature extraction:
  Select a pre-trained model $P$ from the set {VGG16, ResNet, AlexNet}.
  Use the weights $W_p$ of model $P$ to initialize the weights of the corresponding layers in model $M$.
Step 3: Training the CNN model:
for each image $I_i$ in $D_{\text{train}}$ do
     Feed $I_i$ into the CNN model $M$.
     Perform forward propagation to compute the output $O_i = M\left(I_i; W\right)$, where $W$ are the weights of the model.
     Compute the loss $L_i = \mathcal{L}\left(O_i, E_i\right)$, where $L$ is the loss function.
     Perform backpropagation to compute the gradients $\nabla W = \frac{\partial L_i}{\partial W}$.
     Update the weights $W$ using the optimizer $O : W = O(W, \nabla W)$.
end for
  Repeat the process for a defined number of epochs to improve model performance.
Step 4: Combine the CNN model with the multiclass SVM classifier:
  Extract the features $F_i = M_f\left(I_i; W\right)$ from the last layer $M_f$ of the CNN model for each image $I_i$ in $D_{train}$.
  Train a multiclass SVM classifier $S$ using the feature $F_i$ and emotion labels $E_i$.
Step 5: Testing and evaluation:
for each image $I_j$ in $D_{test}$ do
     Feed $I_j$ into the trained model $M$.
     Perform forward propagation to compute the output $O_j = M\left(I_j; W\right)$.
     Predict the emotion label $\hat{E_j} = \arg\max_e O_j(e)$, where, $O_j(e)$ is the output score for emotion $e$.
end for
     Compute the accuracy $A$ of the model as the proportion of correct predictions:
  $A = \frac{1}{M}\sum_{j=1}^{M} 1\left(\hat{E}_j = E_j\right)$, where, $1(x)$ is the indicator function, 1 if $x$ is true and 0 otherwise.
```

Pseudocode for the CNN-SVM integration

```
  Algorithm Facial Expression Recognition_CNN_SVM
Input:
      - Pre-processed training dataset D = (I_i, E_i), i = 1, 2, ..., N
      - Pre-trained model P (e.g., VGG16, ResNet, AlexNet)
Output:
      - A: Accuracy of the facial expression recognition model
Step 1: Initialize CNN Model:
  a. Define CNN model M with layers  L_1, L_2, ..., L_n
  b. Choose a pre-trained model P from VGG16, ResNet, AlexNet
  c. Initialize weights W_p of M using P
Step 2: Train CNN Model:
  for each epoch in EPOCHS do:
    for each image I_i in D do:
      a. Forward pass:
        - Compute output O_i = CNN(I_i, W)
      b. Compute loss:
        - L = Loss_Function(O_i, E_i) // (e.g., cross-entropy loss)
      c. Backpropagation:
        - Compute gradients $\nabla W$
        - Update weights W using optimizer (e.g., Adam, SGD)
Step 3: Extract Features for SVM:
  for each image I_i in D do:
      a. Extract feature vector F_i from the last CNN layer
Step 4: Train Multiclass SVM:
  - Train SVM classifier S using (F_i, E_i) pairs
Step 5: Testing and Evaluation:
  for each test image I_j in test dataset D_test do:
    a. Extract feature vector F_j using trained CNN model M
    b. Predict emotion label:
      - E_pred_j = argmax(S(F_j))
  Compute accuracy:
      - A = (1/N) * sum(Indicator(E_pred_j == E_j)) for all test samples
Return A
```

3.3.1 Face detection, recognition, and verification

Face detection, which identifies faces in images and extracts key features, is integral to the proposed methodology. This process distinguishes between face and non-face classes, given that human facial features like mouth, forehead, eyes, and nose bear considerable similarities.

Face recognition involves comparing a given person to a database of individuals to yield a prioritized list of matches. In contrast, face verification validates a specific person's identity against their claimed identity, yielding a binary response. Facial expressions play a crucial role in interpersonal communication, often symbolizing the movements and positions of facial muscles.

3.3.2 Model evaluation

The dataset divided into training and testing sets provides the basis for the emotion recognition carried out by the CNN algorithm. The model's effectiveness was measured in terms of its accuracy, derived from the outcomes of the testing phase. This assessment facilitates necessary modifications, fostering the model's precision and reliability.

The organization and clarity of the methodology ensure an optimal flow, reducing the possibility of misinterpretation or confusion. By maintaining coherence with figures and tables, the text reinforces the descriptions, enabling better comprehension of the methodology. The captions for figures and tables align with their content, maintaining consistency and precision throughout the text.

This study benefits from transfer learning, which allows the model to function well with less training data, in contrast to earlier research, which frequently suffers from overfitting and poor generalization on smaller datasets. MATLAB was used to build the strategy, which served to streamline feature extraction, multiclass classification, and CNN model training and fine-tuning. The comprehensive DL and ML toolboxes in MATLAB offered a productive foundation for creating and assessing models.

4 Results and Discussion

The image processing techniques employed in this study consider the image a two-dimensional signal, with time as an additional dimension when processing images as three-dimensional signals. The recommended methodology uses MATLAB, a fourth-generation programming language, and a multi-paradigm mathematical computing environment. Figure 5 illustrates the image pre-processing steps, including the input images used for analysis, the detected faces, and the resized image. Figure 6 displays the extracted image features.

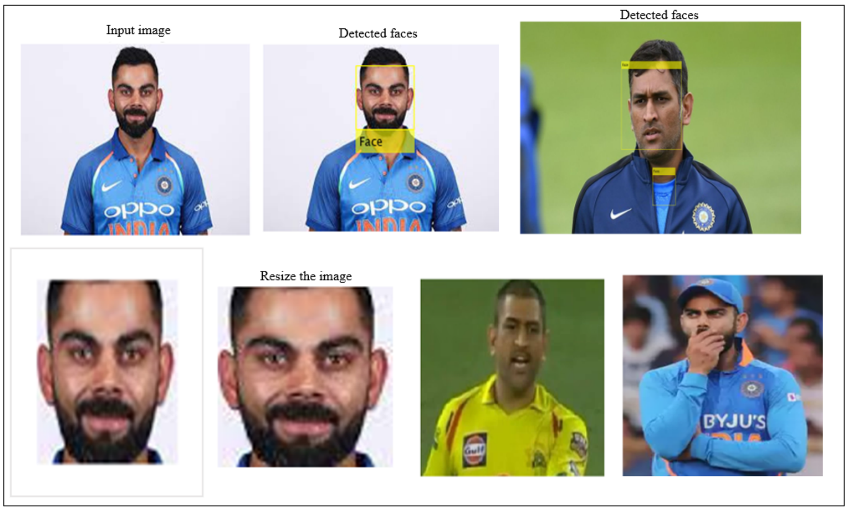


Figure 5. Image pre-processing

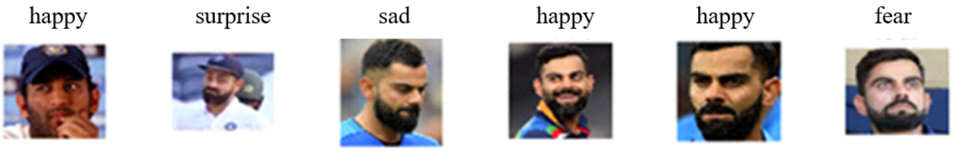
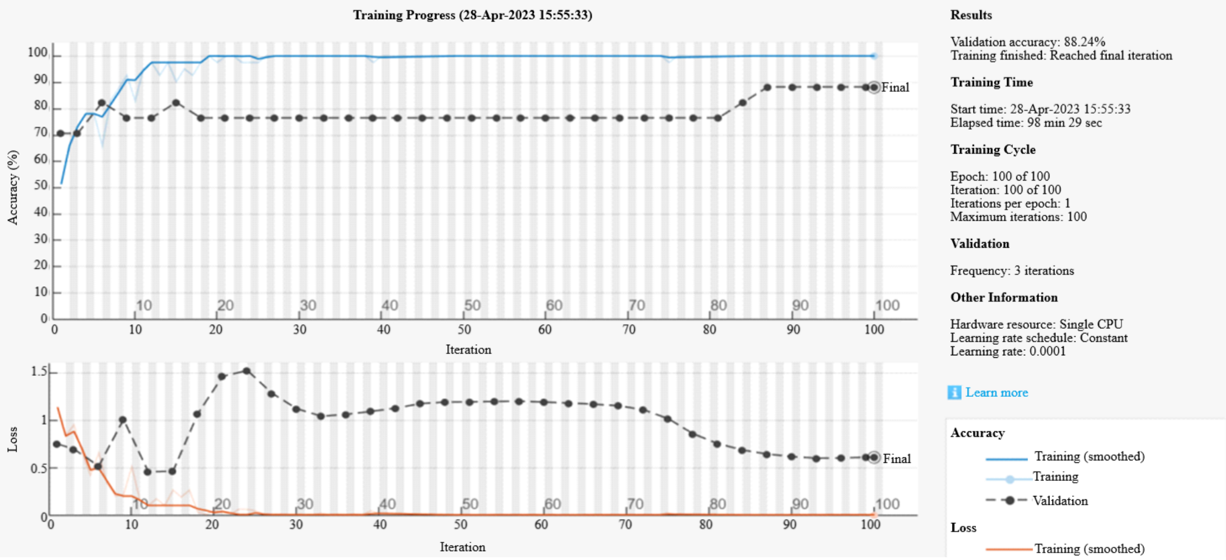


Figure 6. Image features

4.1 Result Analysis for VGG16

The dataset used for training and validation consists of 1,050 face photos with 30 labels, with 725 images used for training and 325 images for validation. The VGG16 model was trained using a gradient descent optimizer for 50 epochs. After approximately 10 epochs, the training accuracy stabilized at 0.98 with a loss of 0.1, indicating no overfitting. The trained VGG16 model achieved an accuracy of 84.52% on the test dataset.

Figure 7 illustrates the training analysis of the VGG16 model, while Figure 8 presents the confusion matrix for VGG16. Figure 9 depicts the confusion matrix for the seven emotional classes using the VGG16 transfer learning model. Figure 10 showcases the predicted classes, and Table 2 provides an overview of the predictive outcomes for the VGG16 model.



Results

Validation accuracy: 88.24%
Training finished: Reached final iteration

Training Time

Start time: 28-Apr-2023 15:55:33
Elapsed time: 98 min 29 sec

Training Cycle

Epoch: 100 of 100
Iteration: 100 of 100
Iterations per epoch: 1
Maximum iterations: 100

Validation

Frequency: 3 iterations

Other Information

Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.0001

Learn more

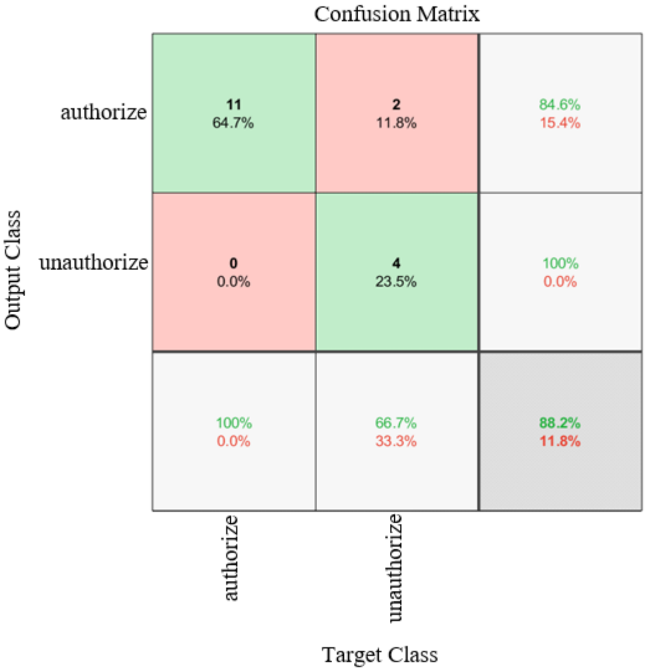
Accuracy

Training (smoothed)
Training
Validation

Loss

Training (smoothed)

Figure 7. GG16 training analysis



	authorize	unauthorize	
authorize	11 64.7%	2 11.8%	84.6% 15.4%
unauthorize	0 0.0%	4 23.5%	100% 0.0%
	100% 0.0%	66.7% 33.3%	88.2% 11.8%

Figure 8. Confusion matrix for VGG16

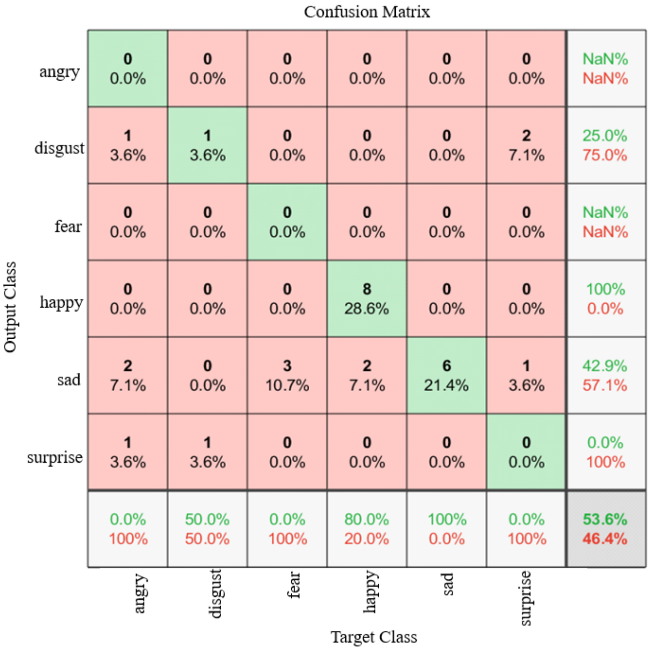


Figure 9. Confusion matrix for six emotional classes using VGG -16

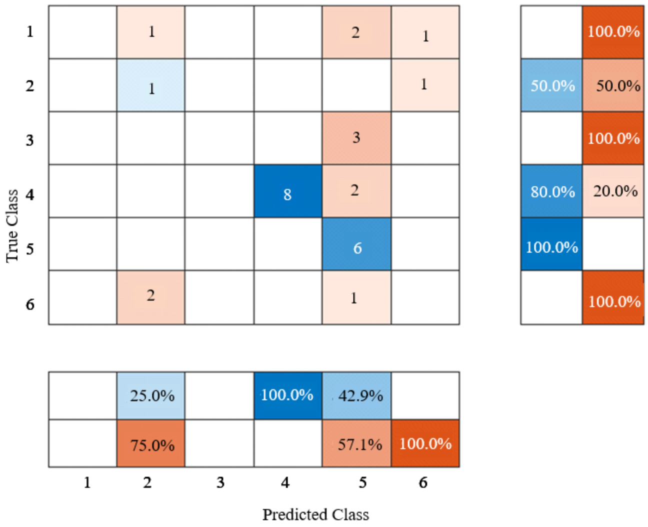


Figure 10. Class prediction

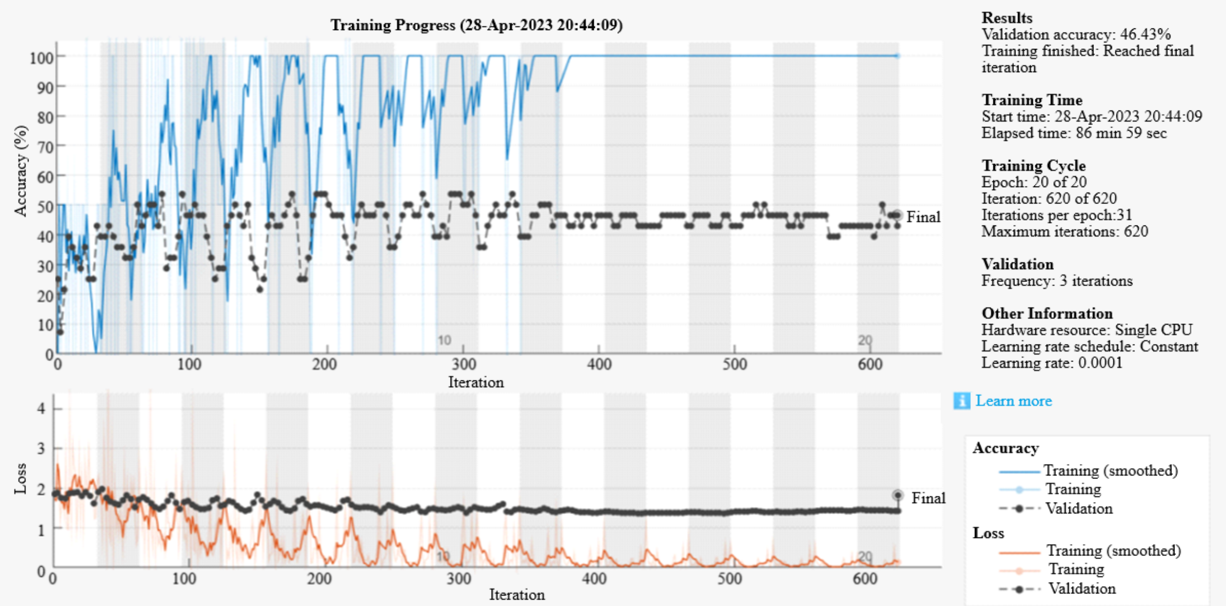
Table 2. Predictive results

Performance Matrices	Performance Value
Accuracy	84.5238
Error	46.4286
Sensitivity	38.3333
Specificity	90.6830
Precision	NaN
False rate	9.3170
F-score	NaN

4.2 ResNet Analysis

A ResNet50 transfer learning model was utilized, training the DCNN with ReLU activations after each convolution layer. The network was trained on the entire dataset using a 58,000-way classifying layer with crystal loss. The batch size was set to 128, and the learning rate was decreased by 0.2 after every 100 iterations, starting from an initial rate of 0.0001.

Figure 11 presents the training analysis of the ResNet50 model, while Figure 12 showcases the confusion matrix for ResNet50. Figure 13 displays the confusion matrix for the seven emotional classes using the ResNet transfer learning model. Feature-classified images using ResNet50 are shown in Figure 14. Table 3 provides the predicted results for the ResNet50 model.



Results
Validation accuracy: 46.43%
Training finished: Reached final iteration

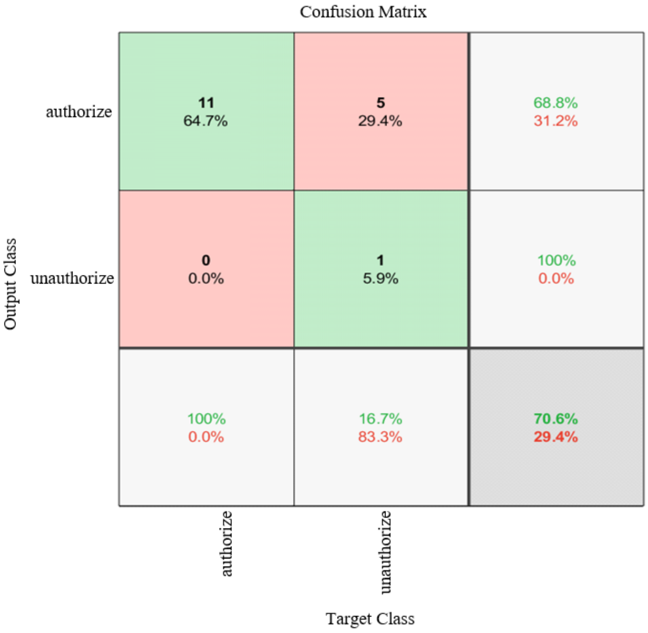
Training Time
Start time: 28-Apr-2023 20:44:09
Elapsed time: 86 min 59 sec

Training Cycle
Epoch: 20 of 20
Iteration: 620 of 620
Iterations per epoch:31
Maximum iterations: 620

Validation
Frequency: 3 iterations

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.0001

Figure 11. Training analysis of ResNet50



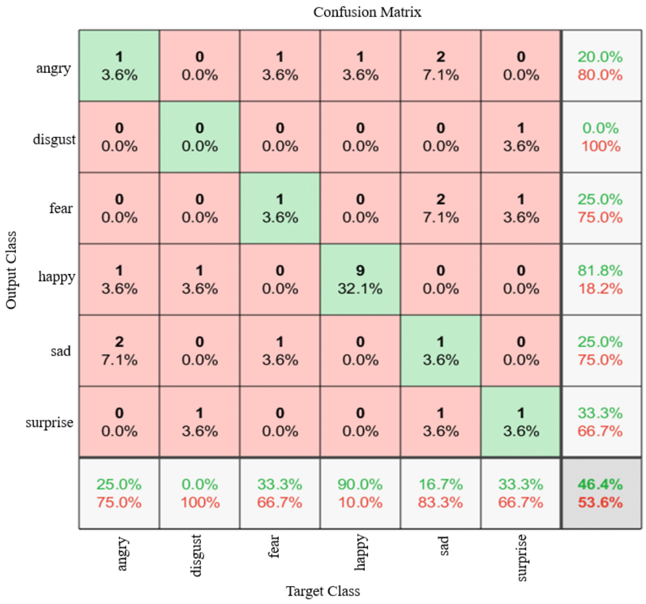
	authorize	unauthorize	
authorize	11 64.7%	5 29.4%	68.8% 31.2%
unauthorize	0 0.0%	1 5.9%	100% 0.0%
	100% 0.0%	16.7% 83.3%	70.6% 29.4%

Figure 12. Confusion matrix of ResNet50

4.3 AlexNet Transfer Learning Analysis

The optimized AlexNet CNN was trained using 1,050 images from each class, with multiple central processing units (CPUs) utilized for training. To prevent memory issues and speed up training, the minimum batch size was set to 1. Training was stopped after 100 iterations as the validation accuracy stabilized. A minimum learning rate of

0.0001 was chosen. Figure 15 illustrates the training analysis of the AlexNet transfer learning model, and Figure 16 presents the confusion matrix for AlexNet.



	angry	disgust	fear	happy	sad	surprise	
angry	1 3.6%	0 0.0%	1 3.6%	1 3.6%	2 7.1%	0 0.0%	20.0% 80.0%
disgust	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 3.6%	0.0% 100%
fear	0 0.0%	0 0.0%	1 3.6%	0 0.0%	2 7.1%	1 3.6%	25.0% 75.0%
happy	1 3.6%	1 3.6%	0 0.0%	9 32.1%	0 0.0%	0 0.0%	81.8% 18.2%
sad	2 7.1%	0 0.0%	1 3.6%	0 0.0%	1 3.6%	0 0.0%	25.0% 75.0%
surprise	0 0.0%	1 3.6%	0 0.0%	0 0.0%	1 3.6%	1 3.6%	33.3% 66.7%
	25.0% 75.0%	0.0% 100%	33.3% 66.7%	90.0% 10.0%	16.7% 83.3%	33.3% 66.7%	46.4% 53.6%

Figure 13. Confusion matrix for six emotional classes using ResNet

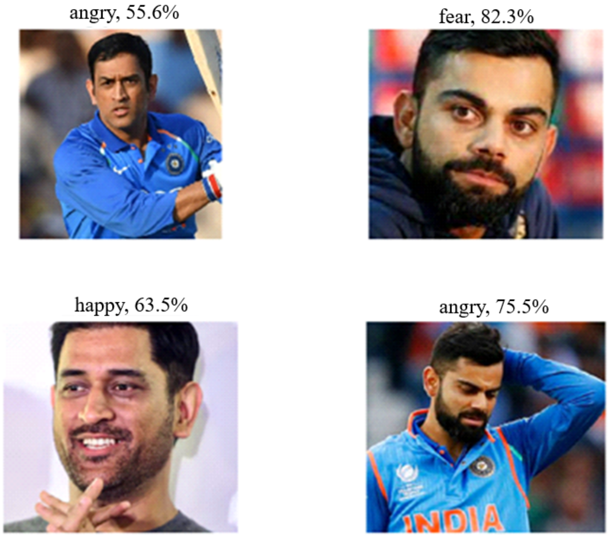
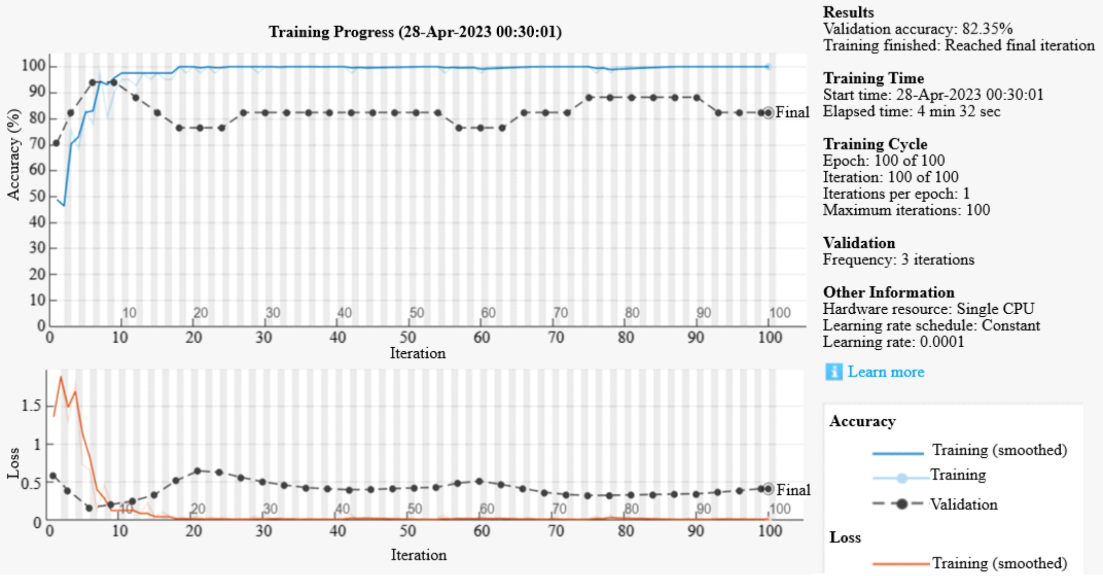


Figure 14. Feature classified images using ResNet

Table 3. Predicted result for ResNet50

Performance Matrices	Performance Value
Accuracy	82.1429
Error	53.5714
Sensitivity	33.0556
Specificity	89.1233
Precision	30.8586
False rate	10.8767
F-score	NaN



Results
Validation accuracy: 82.35%
Training finished: Reached final iteration

Training Time
Start time: 28-Apr-2023 00:30:01
Elapsed time: 4 min 32 sec

Training Cycle
Epoch: 100 of 100
Iteration: 100 of 100
Iterations per epoch: 1
Maximum iterations: 100

Validation
Frequency: 3 iterations

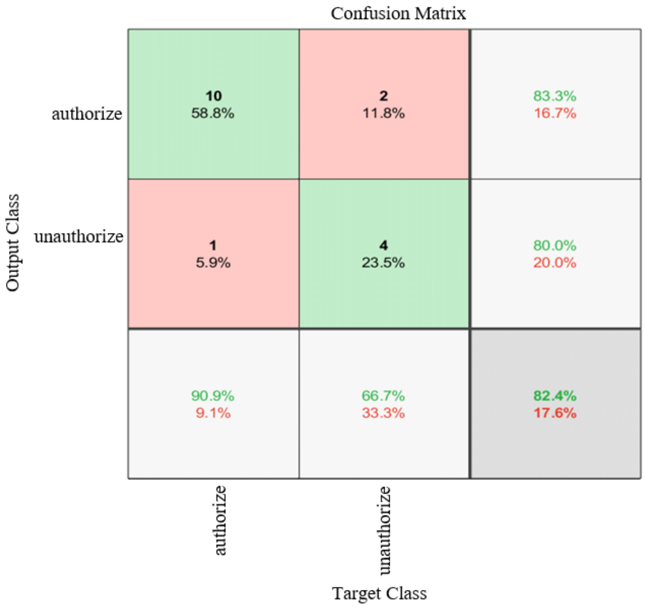
Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.0001

Learn more

Accuracy
Training (smoothed)
Training
Validation

Loss
Training (smoothed)

Figure 15. Training analysis using AlexNet



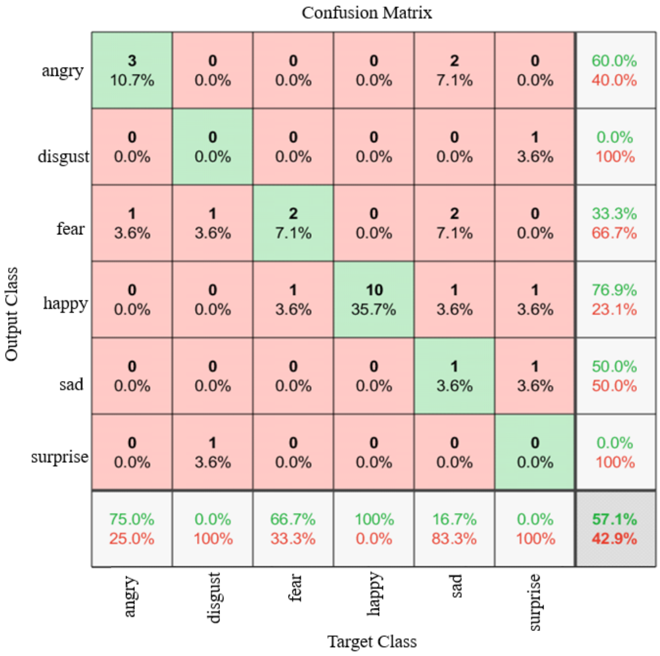
	authorize	unauthorize	
authorize	10 58.8%	2 11.8%	83.3% 16.7%
unauthorize	1 5.9%	4 23.5%	80.0% 20.0%
	90.9% 9.1%	66.7% 33.3%	82.4% 17.6%

Figure 16. Confusion matrix using AlexNet

Figure 17 showcases the confusion matrix for the seven emotional classes using the AlexNet transfer learning model. Figure 18 illustrates the training process of the CNN for image classification, and Figure 19 displays the final classified output images.

Based on the analysis of three transfer learning methods, VGG16, ResNet, and AlexNet obtained accuracies of 84.52%, 82.14%, and 82.40%, respectively. Therefore, the VGG16 transfer learning approach outperformed the other models in terms of accuracy. The AlexNet transfer learning model's overall time consumption was also more efficient than the other transfer models.

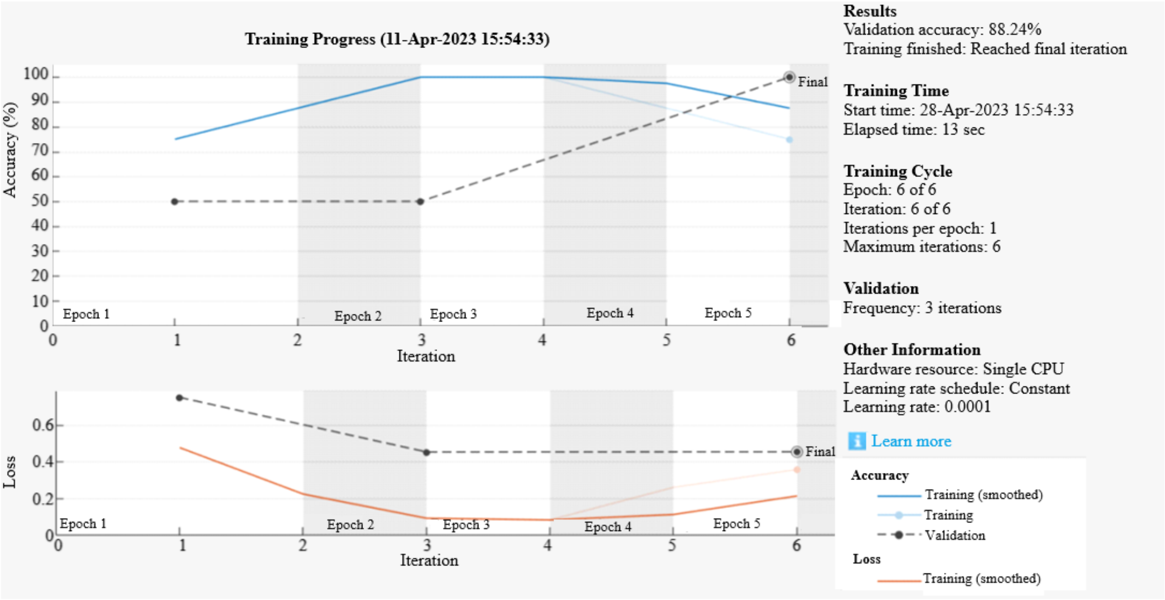
The resilience of the CNN-SVM is a result of the SVM's margin-based decision limits, which efficiently manage class overlap and the CNN's capacity to extract hierarchical, complicated features. On the other hand, the standard SVM's reliance on manually created features hampers its capacity to capture non-linear correlations, while the CNN-alone model's reliance on fully linked layers makes it prone to overfitting, especially with sparse training data. In comparison to previous models, the CNN-SVM achieves a lower accuracy standard deviation and shows superior generalization across several datasets. The CNN-SVM approach's dependability and usefulness for FER tasks are highlighted by this investigation.



Confusion Matrix

	angry	disgust	fear	happy	sad	surprise	
angry	3 10.7%	0 0.0%	0 0.0%	0 0.0%	2 7.1%	0 0.0%	60.0% 40.0%
disgust	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 3.6%	0.0% 100%
fear	1 3.6%	1 3.6%	2 7.1%	0 0.0%	2 7.1%	0 0.0%	33.3% 66.7%
happy	0 0.0%	0 0.0%	1 3.6%	10 35.7%	1 3.6%	1 3.6%	76.9% 23.1%
sad	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 3.6%	1 3.6%	50.0% 50.0%
surprise	0 0.0%	1 3.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0.0% 100%
	75.0% 25.0%	0.0% 100%	66.7% 33.3%	100% 0.0%	16.7% 83.3%	0.0% 100%	57.1% 42.9%

Figure 17. Confusion matrix for six emotional classes using AlexNet



Results
Validation accuracy: 88.24%
Training finished: Reached final iteration

Training Time
Start time: 28-Apr-2023 15:54:33
Elapsed time: 13 sec

Training Cycle
Epoch: 6 of 6
Iteration: 6 of 6
Iterations per epoch: 1
Maximum iterations: 6

Validation
Frequency: 3 iterations

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.0001

Figure 18. CNN network training

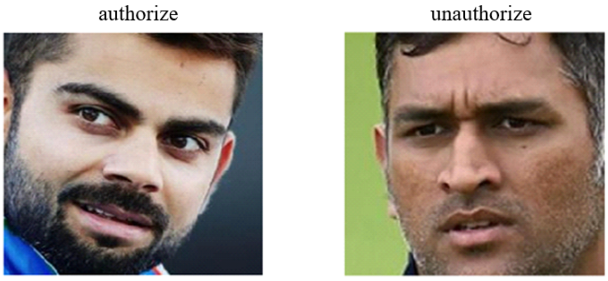


Figure 19. Classified outputs

5 Conclusion

Facial recognition and FER technologies have seen significant advancements due to developments in DL models like ResNet50, VGG16, AlexNet, and CNNs. These technologies have tremendous potential to revolutionize various fields and applications with improved accuracy and dependability. While these technologies hold excellent promise, obstacles, such as more varied and representative training data, privacy issues, and potential bias and discrimination, must be overcome. The suggested technique demonstrates that VGG16 performs better in accuracy while AlexNet performs better in time consumption. With the ongoing improvements in ML and computer vision algorithms, facial recognition and recognition of facial expressions seem to have a promising future. In the future, the development of more advanced DL models is expected, which can further enhance the capabilities of these technologies and provide better solutions that can make them more accessible to various industries. Additionally, greater efforts toward addressing data privacy, ethics, and fairness concerns are expected to ensure that the technology is deployed and used for beneficial purposes. The study shows that combining CNNs and SVMs greatly enhances the performance of FER, with an accuracy of 92% as opposed to 88% for CNN alone and 80% for conventional SVM. This enhancement demonstrates how well CNNs' reliable feature extraction and SVMs' accurate classification work together. With fewer standard deviations in performance measurements, the suggested method also demonstrates improved generalization across datasets and decreased overfitting. It is also computationally efficient by using transfer learning to reduce the amount of time and resources needed for training. Practically speaking, the system's high accuracy and dependability make it appropriate for use in human-computer interaction, allowing for more sensitive and responsive interfaces, and in healthcare, such as stress detection or mental health monitoring. To sum up, the suggested method for FER in conjunction with an SVM classifier and transfer learning with DCNNs, such as VGG16, ResNet, and AlexNet, has important real-world applications. The approach is quite flexible for a range of domains due to its improved accuracy and capacity to generalize across various datasets. The approach could improve the user experience in virtual assistants, video games, and educational aids by enabling emotionally intelligent interfaces that react to users' facial emotions in human-computer interaction. Particularly in telemedicine, where real-time emotion detection can guide improved treatment plans for disorders like anxiety or depression, the technology could help detect emotional distress and offer opportunities for early intervention.

Data Availability

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] R. Ranjan, A. Bansal, J. X. Zheng, H. Y. Xu, J. Gleason, B. Lu, A. Nanduri, J. C. Chen, C. D. Castillo, and R. Chellappa, "A fast and accurate system for face detection, identification, and verification," *IEEE Trans. Biom. Behav. Ident. Sci.*, vol. 1, no. 2, pp. 82–96, 2019. https://doi.org/10.1109/TBIOM.2019.2908436

[2] S. Petrovica, A. Anohina-Naumeca, and H. K. Ekenel, "Emotion recognition in affective tutoring systems: Collection of ground-truth data," *Procedia Comput. Sci.*, vol. 104, pp. 437–444, 2017. https://doi.org/10.1016/j.procs.2017.01.157

[3] J. H. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, 2020. https://doi.org/10.1016/j.inffus.2020.01.011

[4] S. L. Peng, H. Y. Jiang, H. X. Wang, H. Alwageed, Y. Zhou, M. M. Sebdani, and Y. D. Yao, "Modulation classification based on signal constellation diagrams and deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 718–727, 2019. https://doi.org/10.1109/TNNLS.2018.2850703

[5] M. Mohammadpour, H. Khaliliardali, S. M. R. Hashemi, and M. M. AlyanNezhadi, "Facial emotion recognition using deep convolutional networks," in *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, Tehran, Iran, 2017, pp. 17–21. https://doi.org/10.1109/KBEI.2017.8324974

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017. https://doi.org/10.1145/3065386

[7] R. X. Cui, M. Y. Liu, and M. H. Liu, "Facial expression recognition based on ensemble of multiple CNNs," in *11th Chinese Conference on Biometric Recognition (CCBR 2016)*, Chengdu, China, 2016, pp. 511–518. https://doi.org/10.1007/978-3-319-46654-5_56

[8] M. F. Valstar, M. Mehu, B. H. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. Syst. Man Cybern. B*, vol. 42, no. 4, pp. 966–979, 2012. https://doi.org/10.1109/TSMCB.2012.2200675

[9] H. Jung, S. Lee, S. Park, B. Kim, J. Kim, I. Lee, and C. Ahn, "Development of deep learning-based facial expression recognition system," in *2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, Mokpo, Korea (South), 2015, pp. 1–4. https://doi.org/10.1109/FCV.2015.7103729

[10] S. G. Liu, H. X. Wang, and M. Pei, "Facial-expression-aware emotional color transfer based on convolutional neural network," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 1, pp. 1–19, 2022. https://doi.org/10.1145/3464382

[11] P. Barros, N. Churamani, and A. Sciutti, "The FaceChannel: A fast and furious deep neural network for facial expression recognition," *SN Comput. Sci.*, vol. 1, p. 321, 2020. https://doi.org/10.1007/s42979-020-00325-6

[12] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, 2021. https://doi.org/10.3390/electronics10091036

[13] A. A. Almisreb, N. Jamil, and N. M. Din, "Utilizing AlexNet deep transfer learning for ear recognition," in *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, Kota Kinabalu, Malaysia, 2018, pp. 1–5. https://doi.org/10.1109/INFRKM.2018.8464769

[14] S. Khan, E. Ahmed, M. H. Javed, S. A. Shah, and S. U. Ali, "Transfer learning of a neural network using deep learning to perform face recognition," in *2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, Swat, Pakistan, 2019, pp. 1–5. https://doi.org/10.1109/ICECCE47252.2019.8940754

[15] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "FaceQnet: Quality assessment for face recognition based on deep learning," in *2019 International Conference on Biometrics (ICB)*, Crete, Greece, 2019, pp. 1–8. https://doi.org/10.1109/ICB45273.2019.8987255

[16] K. Kalpana, C. Tharani, and B. Paulchamy, "FPGA implementation of noise removal images using modified trimmed median filter," *Int. J. Sci. Res.*, vol. 3, no. 12, pp. 1779–1783, 2014. https://www.ijsr.net/getabstract.php?paperid=SUB14826

[17] R. M. Prakash, N. Thenmoezhi, and M. Gayathri, "Face recognition with convolutional neural network and transfer learning," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2019, pp. 861–864. https://doi.org/10.1109/ICSSIT46314.2019.8987899

[18] V. V. Teresa, J. Dhanasekar, V. Gurunathan, and T. Sathiyapriya, "An efficient technique for image compression and quality retrieval in diagnosis of brain tumour hyper spectral image," in *Machine Learning and Deep Learning Techniques for Medical Science*. CRC Press, 2022, pp. 27–44. https://doi.org/10.1201/9781003217497-2

[19] G. B. Huang, H. M. Zhou, X. J. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst. Man Cybern. B*, vol. 42, no. 2, pp. 513–529, 2011. https://doi.org/10.1109/TSMCB.2011.2168604

[20] E. Pranav, S. Kamal, C. Satheesh Chandran, and M. H. Supriya, "Facial emotion recognition using deep convolutional neural network," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2020, pp. 317–320. https://doi.org/10.1109/ICACCS48705.2020.9074302