IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 September 2023, accepted 11 October 2023, date of publication 16 October 2023, date of current version 31 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325108

RESEARCH ARTICLE

Multi-Stream Deep Convolution Neural Network With Ensemble Learning for Facial Micro-Expression Recognition

GULNAZ PERVEEN[1], SYED FAROOQ ALI[1], JAMEEL AHMAD[1], SANA SHAHAB[2], MUHAMMAD ADNAN[1], MOHD ANJUM[3], AND IKRAMULLAH KHOSA[4], (Member, IEEE)

[1]School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan
[2]Department of Business Administration, College of Business Administration, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[3]Department of Computer Engineering, Aligarh Muslim University, Aligarh 202002, India
[4]Department of Electrical and Computer Engineering, COMSATS University Islamabad, Lahore Campus, Islamabad 54000, Pakistan

Corresponding author: Syed Farooq Ali (farooq.ali@umt.edu.pk)

This work was supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, through Researchers Supporting Project PNURSP2023R259.

ABSTRACT Micro-expression recognition has gained much attention in research communities. Among its proposed solutions, deep learning approaches have shown promising results over the past few years. In this paper, we propose a multi-stream deep convolution neural network with ensemble classification for facial micro-expression recognition. The multi-stream network uses the deep features of a residual network, densely connected convolutional network, and visual geometry group. The features of these aforementioned architectures are extracted from their pooling layers and become very resource-intensive due to their high dimensions. The principal component analysis is applied to these features for their dimensionality reduction. Stacking, an ensemble classification technique, is performed on these deep features with three base learners (random tree, J48, random forest) and a meta learner (random forest). Experiments were performed using publicly available datasets, namely: CASME-II, CASME^2, SMIC, and SAMM. The proposed approach (PA) is compared with twelve approaches. The results show that the PA outperformed the existing approaches in terms of accuracy and time efficiency.

INDEX TERMS Deep learning, ensemble classification, convolution neural networks, face recognition, micro-expression recognition.

I. INTRODUCTION

Expression plays a significant role in human communication. Besides verbal communication, non-verbal cues like human gestures and voice pitch depict, human feelings and give feedback. Likewise, facial expression conceals psychology and mental health, criminal offense, and physical pain. Research on facial expression recognition (FER) engaged scientists of many disciplines working in physiology, acoustic, natural language processing, neuroscience, and computer vision [1]. The achievements in computer vision and artificial intelligence make facial expression recognition plausible to implement the basic sense of vision in humans. Such recognition systems have widespread applications. With the advent of robotics especially humanoid robots, a more robust expression recognition system becomes crucial. Other applications of FER include telecommunication, video games, animations, psychiatry, forensics, crime investigation, fraud detection, automobile safety, surveillance, and educational software. Micro Expressions (ME) and Micro Expression Recognition (MER) are important concepts in the field of automatic facial expression recognition (FER). FER focuses on macro-expressions, which are typical facial expressions that last for a longer duration and are easier to observe and recognize. Micro Expressions are very brief, involuntary facial expressions that occur in a fraction of a second (typically lasting between 1/25th to 1/5th of a second). Micro-expressions are very brief and last for only a fraction

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

of a second. They are unconscious and automatic, often occurring due to emotional suppression or concealment. They often reveal the true underlying emotions of a person, as they are difficult to control consciously. The facial expressions involved in MEs are considered universal across cultures, suggesting that certain emotions have consistent facial manifestations. They involve rapid and subtle movements in specific facial muscles. To consider a video frame as a micro-expression, it should meet the following criteria: short duration, sudden appearance, and emotional intensity.

Several machine learning and AI approaches are used to improve the performance of the FER System. One such approach is Ensemble learning which achieves a top-notch predictive performance with bagging, boosting, and stacking [2]. In ensemble learning, stacking combines heterogeneous base models, organized in one layer, then employs a meta-learner to sum up the outputs of these models. That is a highly effective solution to improve predictive performance with the least effect on time efficiency.

Ensemble learning techniques are popular among researchers over the last decade [3] and they have reported performance improvements using these techniques [4]. These ensemble techniques have been applied in a variety of domains including healthcare, finance, insurance, automobile, manufacturing, bio-informatics, aerospace, cardio-vascular disease, student KPIs measurement, movie review, opinion-making, weather forecast, and many more [5]. In ensemble learning, various classifiers are employed, resulting in better classification accuracy.

Stacking ensemble learning [6] is introduced in [7] that involves the formation of linear combinations of different predictors to give overall improved results. The popularity of ensemble learning inclined our attention to implement it in the FER domain. Stacking ensemble learning methods used in this paper are described as follows. In our FER technique, first, we stacked three individual learners namely J48, Random Forest, and Random Tree that generated a new model. Secondly, the meta-learner combines the results of individual base learners to improve the prediction of our proposed approach (PA) without compromising the time efficiency. In FER systems, overfitting happens due to various factors, namely: the limited size of the training set, the complexity of the classifiers, and the presence of noise [8]. Very little attention was given to ensemble-based learning that reduces over-fitting when the training samples are not enough [9].

In [10], the first-time ensemble method is applied over the three types of features with their corresponding feature extraction methods. Among these aforementioned types, one is the global features extracted using the principal component analysis (PCA) algorithm, the second is the local features by local binary pattern (LBP) algorithm, and the third is the GIST features extracted by the multi-scale and multi-dimensional Gabor filter group. All these three types of features provide different information about the face and hence, affect the face-recognition performance. The ensemble method combines the effect of these three features to improve the FR performance. In the case of micro expression (ME) that last for a few frames, these shallow feature extraction methods do not work well compared to deep learning methods. To deploy a deep-learning approach, the discriminative matrix needs to be derived from the facial images (offset, apex, onset frames) of spontaneous ME videos. The role of the discriminative matrix is to reduce the intra-class distance and increase the inter-class distance simultaneously.

In [11], Gabor deep convolution neural network (GDCNN) a based method is used to capture the discriminative matrix from the facial images, and ensemble learning is used to increase the accuracy of the FER framework. In this work, only GDCNN is used for deep feature extraction in different environments and used so-called GDCNN ensemble learning. The GDCNN outperforms the simple DCNN when used with ensemble learning. The discriminative matrix derived from shallow and deep methods with large training samples may cause over-fitting and under-fitting problems respectively. Ensemble cascade matrix learning (ECML) is used to provide a compromise between under-fitting and over-fitting problems [12]. In the same work, an EC-RMML method is adopted to further improve the performance.

While advancements have been made in this field, still there are several bottlenecks or challenges that researchers and practitioners face. Limited dataset availability, ambiguity and subjectivity in capturing data, variations in lighting conditions and camera angles, variability in expression intensity, occlusion, and noise, limited feature representation, and real-time processing are among the few challenges that make it difficult to maintain high accuracy. Addressing these bottlenecks requires ongoing research and innovation in areas such as dataset collection, feature representation, algorithmic approaches, and real-time processing. Overcoming these challenges will contribute to the advancement and practical applicability of facial micro-expression recognition systems.

The main contributions of our research work are as follows.

- A multi-stream architecture based on ensemble learning (Stacking) is proposed. To the best of the authors' knowledge, Stacking has never been deployed for micro-expression recognition (MER).
- The PA outperformed the existing architectures on SMIC, CASME-II, CAS(ME)^2, and SAMM data sets in terms of accuracy and time efficiency.
- The PA outperformed state-of-the-art deep neural networks.
- We also presented the variants of the PA.
- The PA uses Stacking that reduces the problem of overfitting.

The main contributions of our research work are as follows.

The remainder of the paper is organized as follows: Section II reviews existing MER methods while Section III provides the detail of various datasets used in this research work. Section IV describes the proposed framework for MER. Section V shows the experimental results and detailed discussion while Section VI concludes the paper and gives future directions.

II. RELATED WORK

Ekman and Friesen introduced a Facial Action Coding System (FACS) based on changes in forty-four Action Units (AUs) independently or simultaneously [13] to identify seven universal expressions of anger, contempt, disgust, fear, happiness, sadness, and surprise as shown in Fig. 1. Later, Ekman developed a micro-expression training tool (METT) to detect ME that lasts in a split-second (1/5 to 1/25) duration. Emotions were communicated prior to verbal communication and are helpful in detecting lies, mental health, and dangerous demeanor even if a person tries to hide the real information. Due to the short duration of ME, it was difficult for a trained human to detect it. Computer systems were trained in the past few decades in facial expression recognition (FER) using various methods including long image sequences using optical flow [14], radial basis function network (RBFN) [15], and multi-layer perceptron [16].

FER is still a challenging problem in the field of psychology, physiology, and computer vision. FER was applied on a well-segmented video containing ME from start to end. Before discussing the FER methods, one should be familiar with the basic characteristics of ME, namely: low intensity, involuntariness, short duration, over-fitting problems, lack of training data, and subtle and significant overlap between basic emotions (compound emotions). To solve the last two problems, local methods were used to divide the facial area into well-known sub-regions. The work [17] included facial muscle actions for better recognition of facial behavior. To track the facial action dynamics in an input video with fifteen facial points, particle filtering was used. Polikovsky et al. [18] adopted the ASM to segment posed faces into 12 sub-regions and detected the facial landmarks. The 3-D gradient orientation histogram descriptor was defined for facial muscle movement. The descriptor could find the correlation between frames. They used high-speed (200 fps) camera to capture at least ten frames for the detection of thirteen MER in posed faces.

The authors [19] further improved this work with twenty-two action units using Gabor-feature-based boosted classifiers combining support vector machine (SVM), GentleBoost, and hidden Markov model (HMM). Twenty facial fiducial points were tracked in a sequence of images using particle filtering. The PA achieved a testing accuracy of 95.3% for AU recognition on a benchmark dataset for facial expressions and 72% for spontaneous expressions.

Brain-inspired neural networks had better results for the classification of ME but they needed large data sets to avoid misclassification and class imbalance problems. Data augmentation techniques were used to increase the number of training samples currently available in public image-labeled databases such as Karolinska Directed Emotional Faces (KDEF) [20] and Extended Cohn-Kanade (CK+) [21]. Apart from these challenges, a lack of cross-database evaluation posed difficult invalidation.

The deep-learning techniques were preferred as they used hierarchical classification of ME through multiple layers. Unlike local methods of FER, the holistic methods considered the whole posed face for the detection of ME. Pfister et al. in [22] used Spatio-temporal holistic features to detect sixty-eight landmarks using the ASM after normalizing the frame number of ME video sequences. Further, these landmarks were used to align the face in case of head movements. Then LBP-TOP has used for ME feature extraction while the classification of ME was performed by SVM, Multiple Kernel Learning (MKL), and Random Forest (RF).

In literature, color space models were presented for the detection of ME. The well-known RGB color space model was transformed linearly or non-linearly into YCbCr, CIELab, and CIELuv models. The transformed perceptual color space models had less mutual information than the basic RGB color space model. Wang et al., in their synonym, works [23], introduced a novel model named as Tensor Independent Color Space (TICS) to recognize ME. TICS learned from samples and treated three color components as much independently as possible. TICS achieved better results compared to RGB, CIELab, and CIELuv using CASME and CASME-2 data repositories. In literature, mostly LBP-TOP and its variants or HOOF were used as a feature vector.

Owing to challenging spontaneous ME recognition, Liu et al. [24] proposed a thirty-six region of interest (ROI) based Main Directional Mean Optical Flow (MDMO) normalized static feature vector that considered the spatial location and local statistic motion information for a very small dimension i.e., 36 × 2=72. The optical flow method aligned all frames of a ME video clip and removed noise due to head movements. Then SVM classifier was used for ME recognition. Three databases such as CASME, CASME-II, and SMIC were used to validate the performance. MDMO features were extracted by averaging a set of atomic features frame by frame in a video clip. However, MDMO might lose the manifold structure inherent in feature space. So an effective dictionary-based learning methodology was adopted [25] to learn from a ME video dataset. A distance metric was used to capture the sparsity of sample points in the feature resulting in a sparse MDMO. The new metric was added to the classic graph regularized sparse coding (GraphSC) scheme. A small amount of training data was required in sparse-MDMO that potentially used unsupervised learning with sparse coding. On several spontaneous ME datasets, sparse MDMO features had shown improvements on other representative features like LBP-TOP, STCLQP, MDMO, and FDM.

Xu et al. [27] used Facial Dynamics Map based on an optical flow estimation algorithm to perform pixel-level

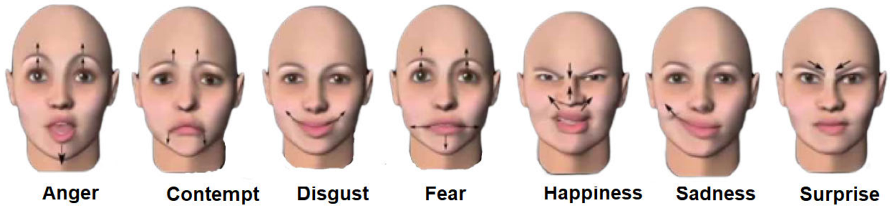


FIGURE 1. Seven micro-expressions.

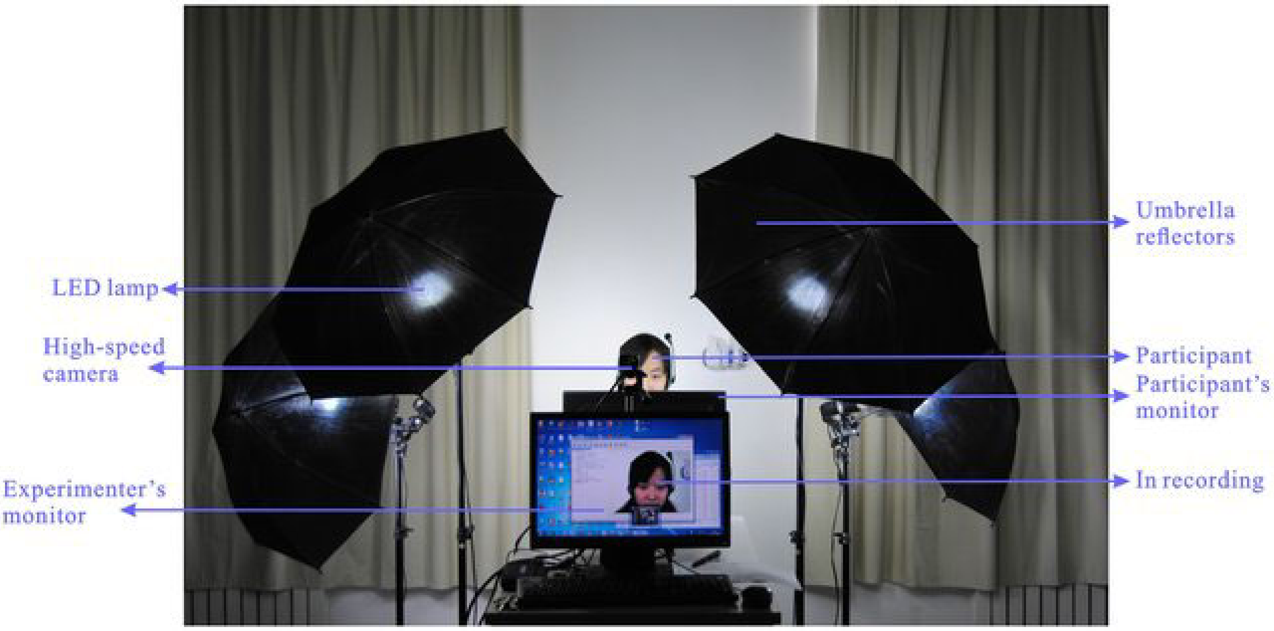


FIGURE 2. Recording and elicitation of micro-expressions [26].

alignment and characterized a ME sequence. Further, the principal optical flow feature was used to track the local movement, and recognition was done by the SVM classifier. In literature, a fixed spatial division grid 8 × 8 was utilized to partition the facial images into a few facial blocks before extracting Spatio-temporal descriptors to distinguish ME. By changing the size of facial blocks the discriminative ability of the descriptor would be affected. Therefore, it was important to find the optimal size of the block [28]. A hierarchical spatial division scheme for Spatio-temporal descriptors was proposed to determine which facial block was suitable for which ME samples. Furthermore, kernelized group sparse learning (KGSL) was modeled to process the descriptors. In [29], authors proposed a deep architecture for facial fiducial points detection and combined motion vectors for ME estimation.

Xie et al. [30] provided a comprehensive review of recent advances in facial ME analysis to discuss new research directions and challenges. In [31], this paper offered a solution to MER using Uniform Local Binary Patterns based on accordion spatiotemporal representation using random forest. The experiments were conducted on the CASME-II dataset yielding an accuracy of 81.38%. Sun et al. [32] proposed an approach for MER based on a multiscale spatiotemporal LBP-TOP descriptor that yielded 77.3% accuracy using the CASME-II dataset. Sadeghi and Raie proposed an approach for MER that was based on a metric learning method to classify histogram data using k-NN [33]. The experiments were conducted on CK+, SFEW, and RAF-DB datasets.

Zhou et al. [34] proposed a CNN, FeatRef, for ME recognition which involved 3 stages of feature refinement, namely: expression-specific feature distilling, expression-shared feature learning, and expression-specific feature fusion. The PA was validated with SMIC-HS, CASME-II and SAMM and claimed accuracy of 70.11%, 89.15% and 73.72% respectively on these datasets.

The previously discussed methods were used for FER from short-duration videos. However, it was proposed in the literature to spot ME from long videos. In [35], spontaneous ME convolutional neural network (SMEConvNet) was the first deep-learning model used for spotting ME from long videos. SMEConvNet comprised four pairs of convolutional and pooling layers to extract 500 features from each frame. The resulting feature matrix is then first processed using difference, squared, and sum operations and at the end

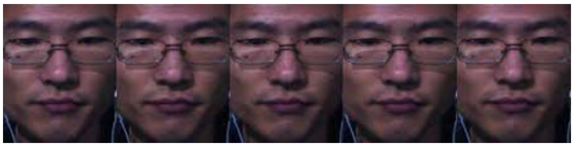


FIGURE 3. Key images of SMIC with negative expression.

manipulated by a sliding window. The sliding window located the largest value and this value found the apex frame. In the literature, a CNN-based micro-FER framework was used for still images. However, spontaneous ME videos needed to exploit both Spatio-temporal information. In [36], 3D-CNN model having two layers, MicroExpSTCNN and MicroExpFuseNet, was used for spontaneous MER. The first layer used the full spatial information and the second layer was used for the fusion of the eyes and mouth region.

In [37], a combined spotting and detecting MEs framework has been presented. In this framework, the temporal interpolation method (TIM) unified the video sequence length, then the peak detection and feature difference contrast method spotted the ME window with onset, apex, and offset frames. Further, the Eulerian video magnification method magnified the motion to overcome low intensity. Finally, three different feature descriptors LBP, HOG, and HIGO were deployed for the ME recognition task. Recurrent Convolutional Neural network (RCNN) has a better ability to describe concealed ME compared to handcrafted features by leveraging the temporal changes.

In [38], RCNN was used for MER from captured normalized images from spontaneous ME video. Temporal jittering was utilized to enrich the training samples to facilitate the learning procedure. In [39], three stream CNN was implemented to incorporate temporal faces, local and entire facial region cues in videos from the temporal stream, static, and local spatial stream respectively for MER. Besides, an algorithm was proposed for apex frame detection.

III. DATASETS

We performed our experiments on the following datasets, namely: SMIC [40], CASME II [26], CAS(ME)^2 [41], and SAMM [42]. Rather than emotion identification, these datasets address the recognition of ME. The acquisition setup for recording ME is shown in Fig. 2. The previous publicly available image labeled databases such as Karolinska Directed Emotional Faces (KDEF) [20] and the Extended Cohn-Kanade (CK+) [21] have not been used in this paper because they have fewer number of training samples. Data augmentation techniques are used to increase the number of training samples but still, they have the difficulty of cross-database validation to overcome these challenges. The details of these four ME databases used in this paper are listed below.

A. SMIC

In [40], a new database for spontaneous ME, SMIC, is used to analyze ME recognition obtained through emotional inducement experiments. This database contains 164 ME video clips of sixteen participants recorded with a high-speed (HS) cameras at a resolution of 100 frames per second and named as SMIC dataset. SMIC database had the SMIC-VIS dataset and SMIC-NIR dataset recorded at normal speed cameras with 25 frames per second for both near-infrared and visual light range respectively. There are some drawbacks to the SMIC database. It is not recorded using FACS coding without neutral sequences i.e., the subject's face does not move before the onset frame. The inducement experiment has protocols that the subject reacts and suppresses the emotions. Subject self-reporting is required in the absence of FACS for the categorization of emotion labels. The flickering of light affects the video quality. The video frames have lower resolution with a facial area of 190 × 230 pixels. SMIC has the information of three ethnicities only and does not represent the entire population. A few samples from the SMIC dataset with negative expressions are shown in Fig. 3.

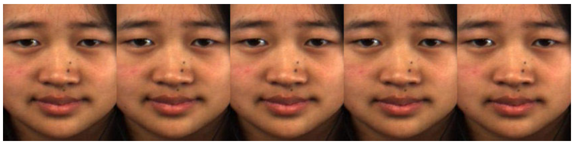


FIGURE 4. Key images of CASME II with happiness expression.

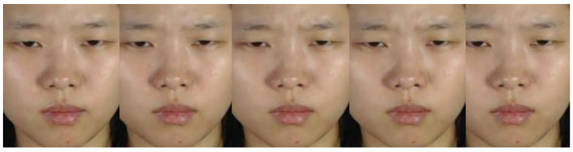


FIGURE 5. Key images of CAS(ME)^2 with anger expression.

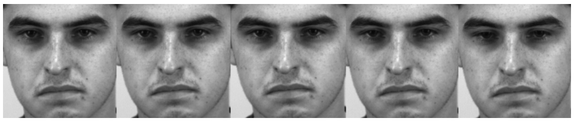


FIGURE 6. Key images of SAMM dataset with anger expression.

B. CASME-II

CASME-II is an extended version of the CASME dataset. Its frame rate is 200 fps and 247 fps with 26 subjects from Chinese ethnicities and it is FACS coded for ME. The facial area in CASME-II has also been increased to 280 × 340 which is larger than CASME and SMIC datasets. A few samples from the CASME-II dataset with happiness expression are shown in Fig. 4.

C. CAS(*ME*)^2

CAS$(ME)^2$ database contains twenty-two subjects with thirteen females and nine males. CAS$(ME)^2$ has 300 macro expressions and 57 micro-expression (ME) sequences. The video resolution is 640 × 480 pixels at 30 fps captured with a camera of 500 ms shutter speed. The image sequences have

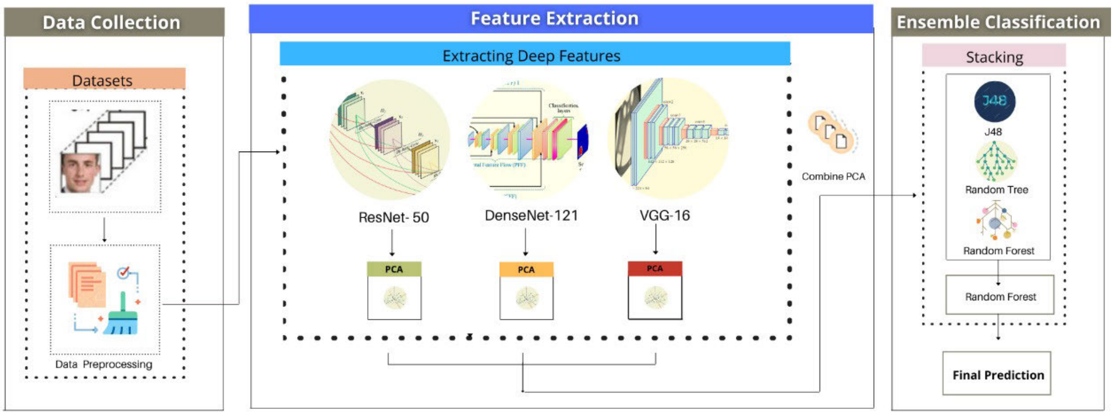


FIGURE 7. Proposed architecture for ME classification.

three classes of anger, happiness, and disgust extracted on the basis of 600 AUs. A few samples from CAS$(ME)^2$ dataset with anger expression are shown in Fig. 5. The main features of this database are summarized below:

- CAS$(ME)^2$ has the advantage of a long video sequence of both macro and micro expressions and it is publicly available. That helps in developing a spotting algorithm in long video sequences.
- All micro and macro-expressions samples are gathered in the same environment over a single subject. That helps research to propose more robust features for discriminating micro and macro expressions.
- For each expression, the dataset collected a combination of facial action units, type of the emotion extracted from the video, and the subject‘s self-reported emotions.

D. SAMM

The SAMM database contains twenty-nine subjects with 159 ME that are divided into 8 ME classes. This data repository contains 159 macro expressions generated using 29 individuals. Fig. 6 shows key images of SAMM. The videos were collected at 200 fps having a resolution of 650 × 960. This high resolution played an important role in revealing the potential micro-movement. SAMM has the highest diversity in its demography among all the ME data sets currently available. This provides a better sample of the population and deals with a variety of ME including real-world scenarios.

IV. METHODOLOGY

In this work, we proposed a multi-stream deep convolution neural network with ensemble learning for facial micro-expression recognition (MER) performing feature enrichment to encode subtle facial changes. We first applied ResNet-50 on the dataset and extracted features from the pooling layer, and the same process was repeated with DenseNet-121 and VGG-16. After extracting features from these architectures, we applied PCA to reduce their dimensions. Stacking was then applied with 3 base learners (J48, Random Forest, Random Tree) and using a meta learner (Random Forest). The images of datasets were pre-processed and resized to 128 × 128. The PA for ME classification is shown in Fig. 7. The PA is further explained with its three deep learning architectural blocks in Section IV-A.

A. DEEP LEARNING ARCHITECTURES

This section discusses various characteristic features of ResNet-50, DenseNet-121, and VGG-16.

1) RESIDUAL NETWORK (RESNET)

To overcome the problem of vanishing gradient and accuracy saturation problem, Kaiming developed ResNet [43], [44], [45]. The deep residual network creates simple stack layers, therefore the ResNet was developed with 34, 50, 101, 152, and even 1202 layers. ResNet-50, a popular deep network, consisted of 1 fully connected layer and 49 convolution layers. MACS and weights for the entire ResNet are 3.9 M and 25.5 M respectively. This architecture is different from the original Convolution Neural Network (CNN) in which each layer feedforward to the next layer [46], [47].The key idea behind ResNet is the use of residual blocks, which enable the network to learn quickly by staying close to the data manifold.

ResNet skips or jumps 2-3 layers to make a biological analogy of pyramidal cells in the cerebral cortex. By skipping layers, the network becomes simple and training happens with fewer layers. Eventually, learning becomes faster with less impact of the vanishing gradient problem. As the network learns the feature space, then it gradually restores the skipped layers. When training diminishes, all layers get expanded. ResNet learns quickly with less risk of overfitting because it remains close to the manifold. If residual parts are not included in a neural network then it traverses larger feature space, due to which, it becomes more vulnerable to perturbations and moves away from the manifold requiring more training data to recover.

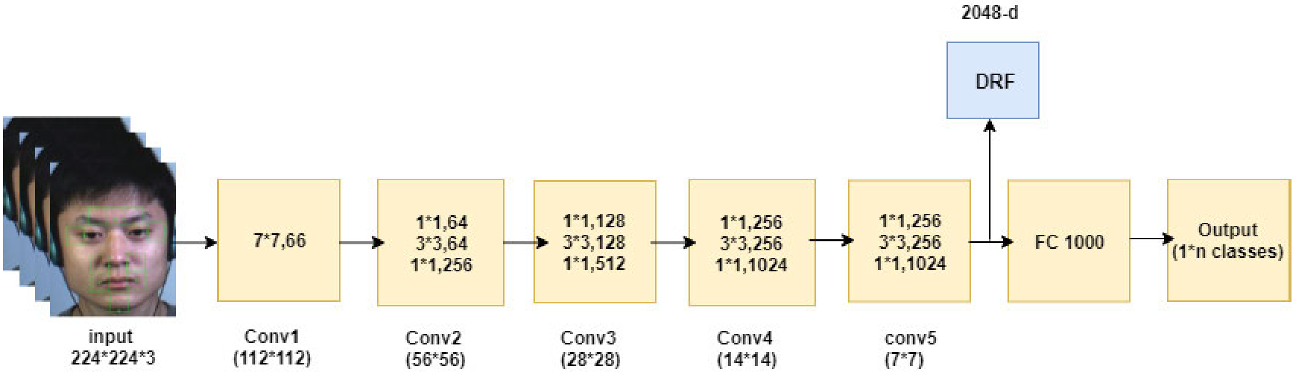


FIGURE 8. ResNet-50 architecture.

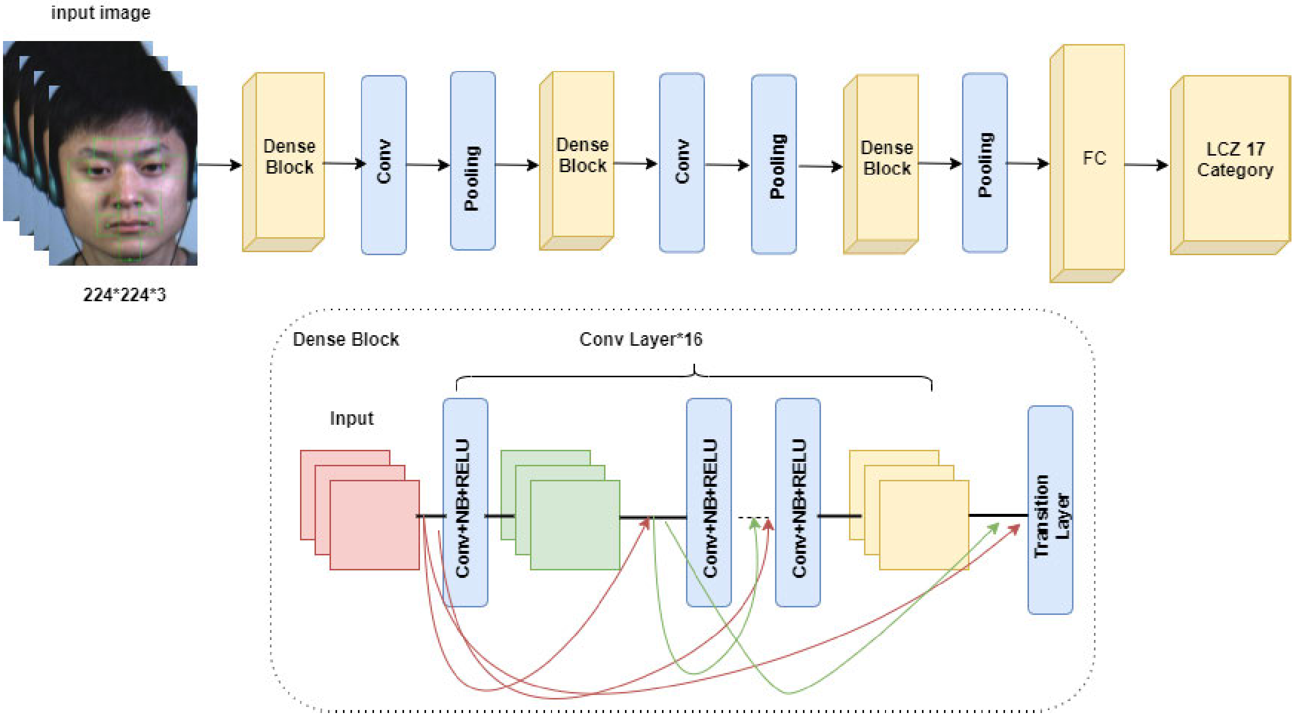


FIGURE 9. DenseNet-121 architecture (top) and its associated dense block(bottom).

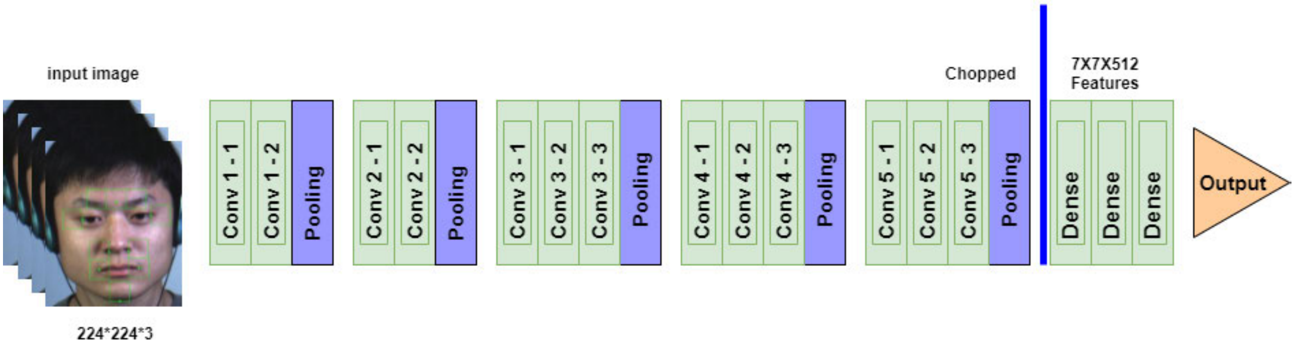
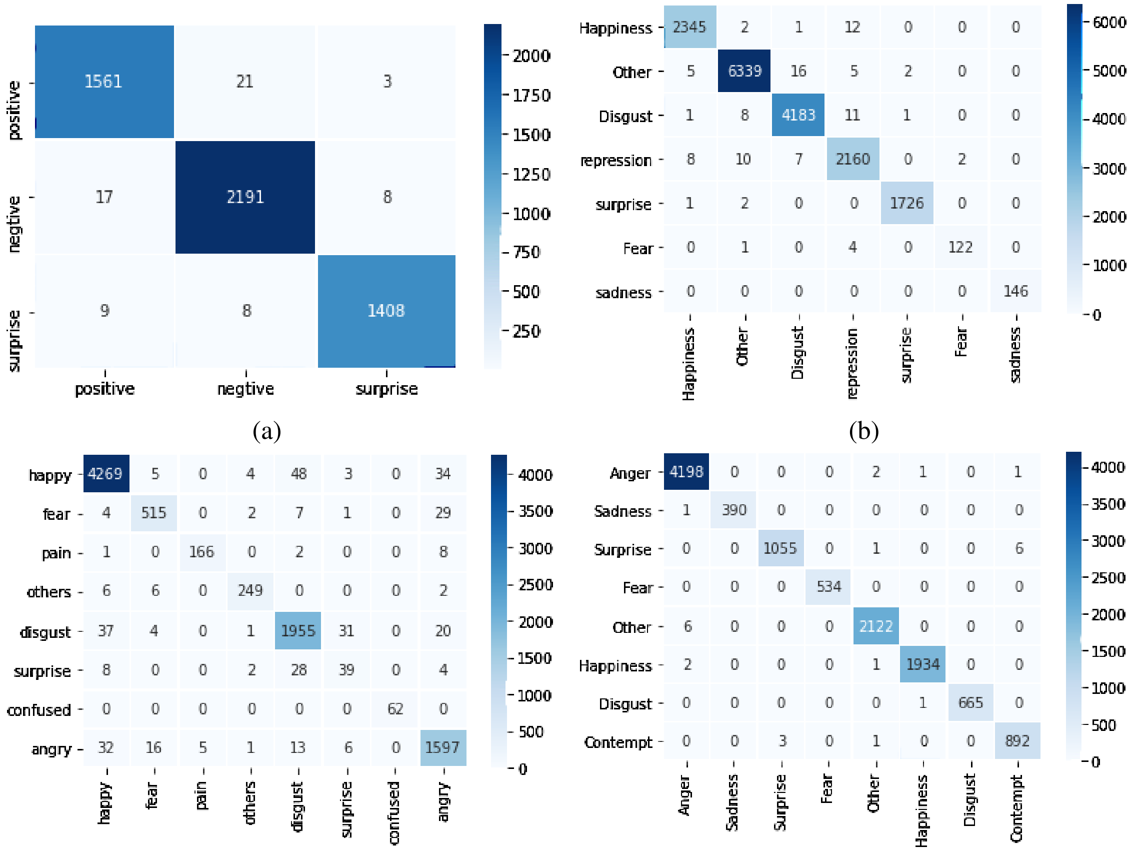


FIGURE 10. VGG-16 architecture.

ResNet is a traditional feedforward network and its output can be determined by the outputs of $(l-1)^{th}$ layer and $f(x_{l-1})$ after carrying out different tasks, such as activation function (ReLu on x_{l-1}) preceded by batch normalization (BN), and convolution with variable filter sizes [48]. Eq.1 defines the x_l i.e., the output of the residual unit. The output equation for



(a)

(b)

(c)

(d)

FIGURE 11. Confusion matrices. Along the x-axis, predicted labels are shown while actual labels exist along the y-axis a) SMIC (3 categories), b) CASME-II (7 categories), c) CAS(ME)^2 (8 categories), d) SAMM (8 categories).

the residual unit (skipped connection) is given below;

$$x_l = f(x_{l-1}) + x_{l-1} \tag{1}$$

The skipped connection simply forwards the input from one layer to the next layer, allowing the gradients to flow directly through the network and alleviating the vanishing gradient problem. The equations for the skipped connection are straightforward as they involve adding the input to the output of a specific layer. The structure of ResNet-50 is shown in Fig. 8.

2) VISUAL GEOMETRY GROUP (VGG)

VGG was the runner-up of the ILSVRC competition 2014 [49]. This work explains the importance of network depth to get better results. VGG contains two convolution layers that use ReLU (activation function) followed by max-pooling and many fully-connected layers. These fully-connected layers also use ReLU. The final layer is the classification layer named as Softmax layer. Three VGG-E models including VGG-19, VGG-16, and VGG-11 were developed having 19, 16, and 11 layers respectively. VGG-19, VGG-16, and VGG-11 have 16, 13, and 8 convolution layers respectively. However, they have the same 3 fully-connected layers. VGG-19 contains 15.5 M MACS and 138 M weights, making it a highly resource-intensive model. Fig. 10 shows the structure of VGG-16.

3) DENSELY-CONNECTED NEURAL NETWORK (DENSENET)

Huang et al. in 2017 [50] developed a convolution neural network in which the output of each layer was connected to every successor layer, forming dense connectivity among layers, and hence it was named DenseNet. DenseNet-121, DenseNet-160, and DenseNet-201 are different versions of DenseNet with 121, 160, and 201 layers respectively. In DenseNet, there exist transition blocks between any two adjacent dense blocks. The feature vector from all the preceding layers is passed as an input to each layer. The l^{th} layer received all the feature maps from previous layers of $x_0, x_1, x_3 \ldots x_{l-1}$ as input which can be modeled by [51] Eq. (2).

$$x_l = H_l[(x_0, x_1, x_2 \ldots x_{l-1})] \tag{2}$$

where $(x_0, x_1, x_2 \ldots x_{l-1})$ are the concatenated features for layers 0 to $l - 1$ and the multiple inputs of $H_l(.)$ are concatenated in Equation (2) into a single tensor. It performs 3 operations: 3 × 3 convolution operation and ReLU

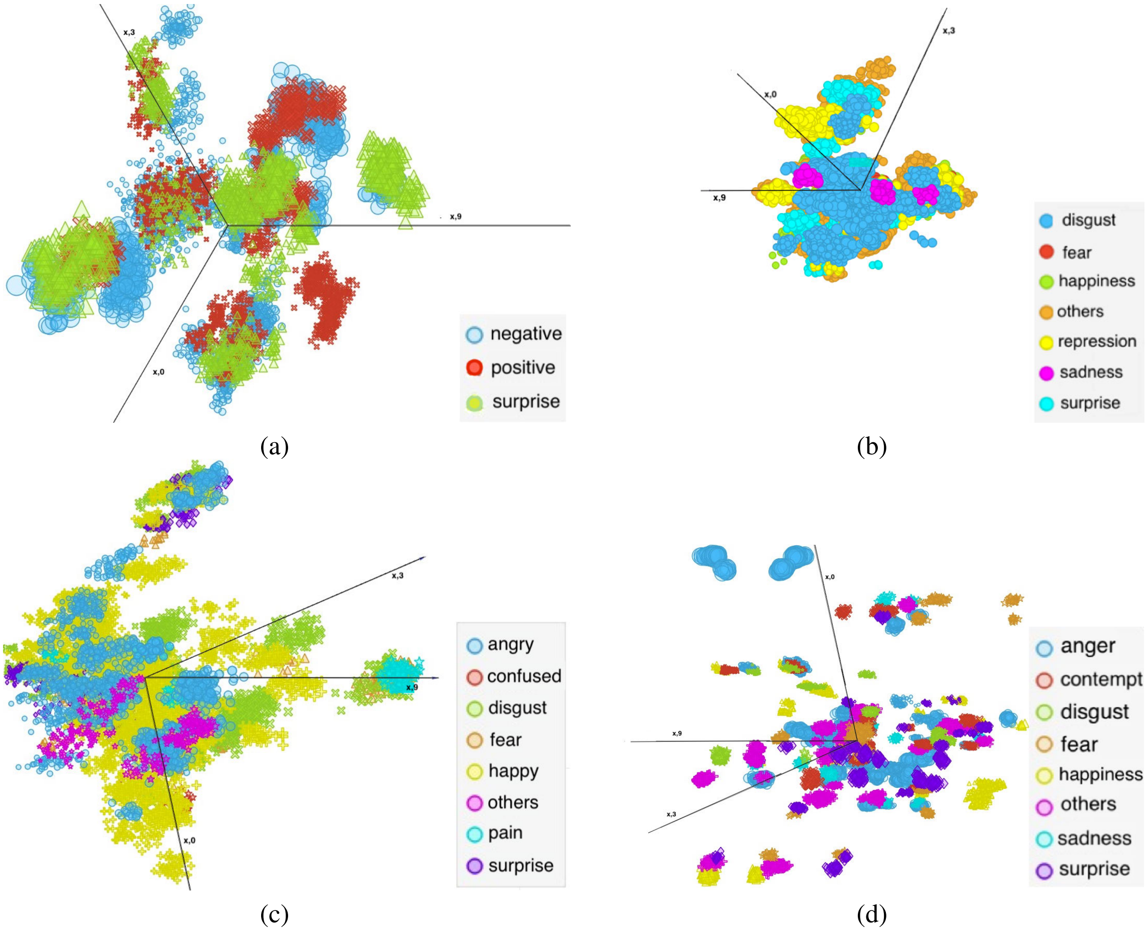


FIGURE 12. Visual representation of features (first three principal components, i.e., A1, A2, and A3) using linear projections of a) SMIC (3 categories), b) CASME-II (7 categories), c) CAS(ME)^2 8 categories), d) SAMM (8 categories).

TABLE 1. Ablation Study: Comparison of PA in terms of percentage accuracy by replacing various classifiers using SMIC, CASME-II, CAS(ME)^2, & SAMM datasets. DT=Decision Tree, RF=Random Forest.

	CASME-II	CAS(ME)^2	SAMM	SMIC
DT	99.24	96.01	98.91	98.66
J48	99.24	95.73	98.72	98.58
RF	95.60	86.98	98.22	92.55
PA	99.39	97.12	99.10	99.82

preceded by BN. The transition layers between the blocks perform convolution and pooling. Hence, the model obtained yielded promising results with an optimal network parameter for tasks related to the recognition of objects. Fig. 9 depicts the structure of DenseNet-121.

B. FEATURES EXTRACTION

We extract deep features from pooling layers of ResNet-50, VGG-16, and DenseNet-121.

After seeing the confusion matrices in Fig. 11, it can be inferred that the percentage errors of our deep features for SMIC, CASME-II, CAS(ME^2), and SAMM are 1.26%, 0.55%, 0.22% and, 4.12% respectively. These low values of mis-classification rate show that our proposed deep features were able to classify the emotions in an effective manner. Similarly, the percentage errors of each category of these afore-mentioned datasets are also very low. For example, in the case of SMIC, the percentage errors of surprise, positive, and negative are 0.33%, 0.48%, and 0.46% respectively. In CASME-II, the percentage error of happiness, other, disgust, repression, surprise, fear, and sadness are 0.09%, 0.16%, 0.12%, 0.16%, 0.02%, 0.03%, and 0% respectively. In CAS(ME^2), the percentage errors of happy, fear, pain, others, disgust, surprise, and confused are 1.02%, 0.46%, 0.12%, 0.15%, 1.01%, 0.57%, 0.00%, and 0.79% respectively. In SAMM, the percentage errors of sadness, anger, fear, surprise, other, happiness, contempt, and disgust are 0.01%, 0.03%, 0.00%, 0.06%, 0.05%, 0.03%, 0.03%, and 0.01% respectively.

1) DIMENSIONALITY REDUCTION

The combined deep features used in the PA has very high dimensions that make them computationally very expensive.

TABLE 2. Optimized hyperparameters of deep architectures of PA. LR = Learning Rate, BS = Batch Size, Opt = Optimizer, AF = Activation Function.

Model	Hyperparameter
DenseNet-121	Epoch: 50 steps per Epoch: 2000 LR: 0.00001 BS: 32 Opt: SGD AF: Relu Image Size: 224*224*3 PCA: 50 Feature Set: 1024
VGG-16	Epoch: 50 steps per Epoch: 2000 LR: 0.00001 BS: 32 Opt: adam AF: Relu Image Size: 224*224*3 PCA: 50 Feature Set: 512
ResNet-50	Epoch: 50 steps per Epoch: 2000 LR: 0.00001 BS: 32 Opt: Adam AF: Relu Image Size: 224*224*3 PCA: 50 Feature Set: 2048

TABLE 3. Optimized hyper-parameters of ensemble classifier (stacking) of PA.

ML Classifier	Hyper-parameter
J48	minimum objects: 1000 MDL correction: True sub tree raising: True
Random Forest	Bootstrap: True max depth: 100 max feature:2 min sample leaf:2 estimators (n): 100 min sample split:10
Decision Table	criterion: gini max depth: none min samples split:2

Hence, PCA was applied for dimensionality reduction [52]. Fig. 12 shows the visual representation of the first three principal components of the proposed deep features for SMIC, CASME-II, CAS(ME)^2, and SAMM datasets respectively.

PCA reduces the dimensionality of high-D datasets without significant loss of information. The newly formed datasets will trade a little bit of accuracy compared to the original one. Machine Learning algorithms run faster over reduced dimensions. In the PA, the first three principal components are used for each data set as shown in Fig. 7 and the deep features are extracted with the minimum possible dimensions.

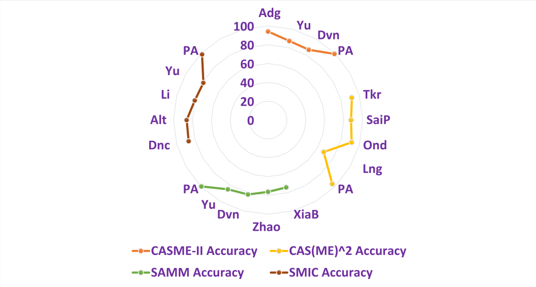


FIGURE 13. Accuracy Comparison of PA with other approaches using SMIC, CASME-II, CASME^2 and SAMM datasets.

C. STACKING

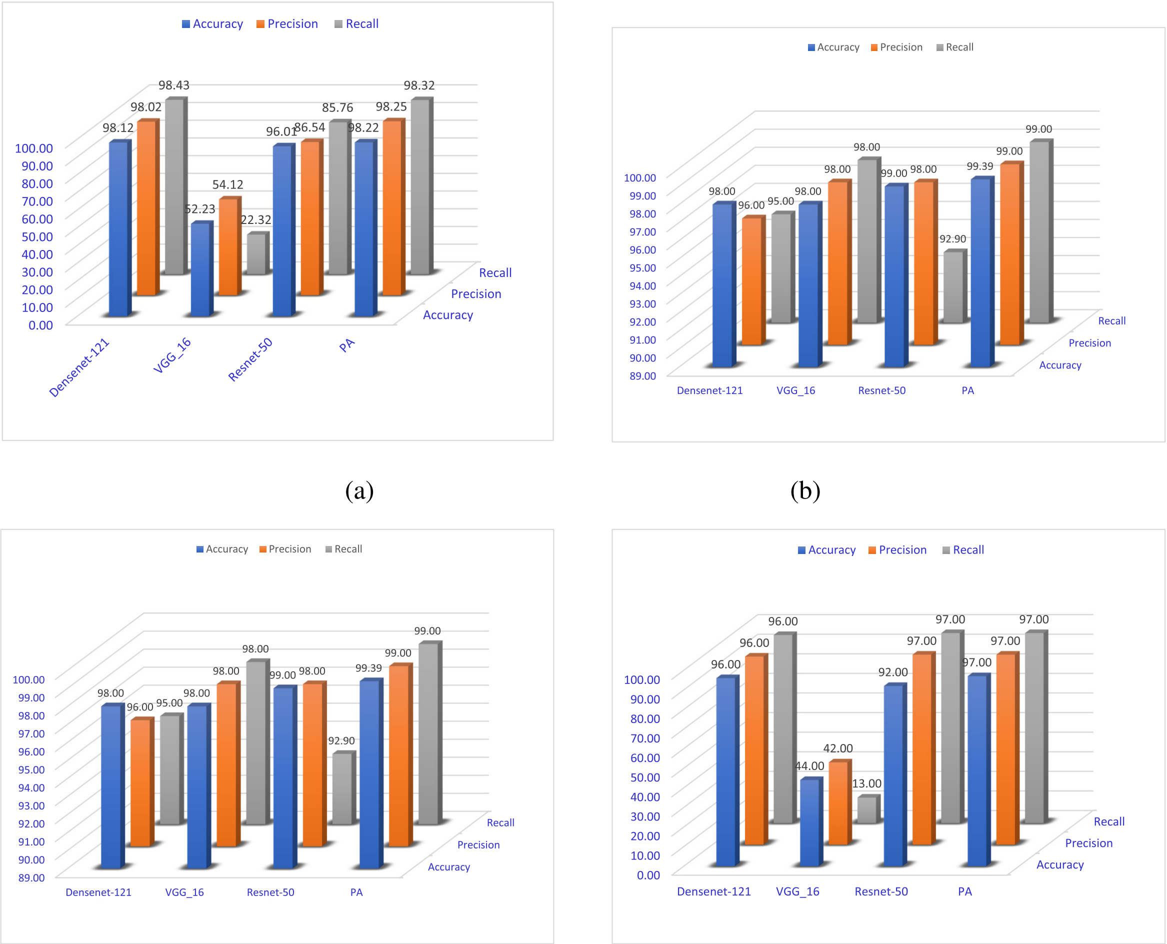
Stacking is one of the ensemble learning algorithms that are appropriate when multiple machine learning algorithms have skills on a data set in different ways. Moreover, the error in predictions of different machine learning algorithms is highly uncorrelated. The multiple machine learning algorithms are called base learners and they hit the problem in different ways. Level-0 models (base models) are fit to the training data and their predictions are compiled. The output label or probability values are generated from base models using k-fold cross-validation. The base learners are trained over the reduced deep features extracted from the pooling layer of deep learning models, reduced through PCA, and then combined. Then Level-1 meta-learner model is used to combine the prediction of the base model. The meta-learner estimate out-of-fold predictions by calculating the mean of each 10-fold cross-validated value. The level-0 (base models) and level-1 (meta learner) formed a super learner machine learning algorithm. The super-learning algorithm's working is explained below.

First, set up the ensemble with a specific list of L base learners and a meta-learner. Secondly, use k-fold cross validation to train L base learners on the training set. All base learners use the same k-fold cross validation and predict $N \times L$ feature matrix represented by Z in Equation (3). Here N is the number of cross-validated values predicted by each base-learner in the list of ensemble L. The feature matrix Z, along with the original response vector (y), is called the "level-one" training data and n is the number of rows in the training data for level-one.

$$n\left\{p_1\right\}\left\{p_2\right\}\ldots.\left\{p_L\right\}\left\{y\right\} \rightarrow \left[Z\right]\left[y\right] \tag{3}$$

Secondly, train the meta-learning algorithm on the level-one data ($y = f(Z)$). In the Third step, the "stacking ensemble model" generates predictions on the test data and that is why it is called ensemble prediction.

In this article, the super learning algorithm is used with three machine learning level-0 base models namely: J48, Random Forest, and Random Tree. Each base learning model uses similar 10- fold cross validation of training data and calculate prediction values and labels. Then level-one data is used by the Random Forest meta-learner algorithm to refine the prediction of the base learner.



(a)

(b)

(c)

(d)

FIGURE 14. Performance of PA with state-of-the-art deep architectures using a) SMIC (3 categories), b) CASME-II (7 categories), c) CAS(ME)^2 (8 categories), d) SAMM (8 categories) datasets.

V. EXPERIMENTS AND RESULTS

The PA was compared with state-of-the-art approaches using standard datasets, namely: SMIC with three categories, CASME II with eight categories, CAS(ME)^2 with seven categories, and SAMM with eight categories. Our PA was also compared with existing state-of-the-art deep architectures (DenseNet-121, VGG-16, and ResNet-50) in terms of accuracy, precision, and recall using the split ratio of 90:10. One of the reasons for the better performance of PA includes stacking with Random Forest.

A. ENVIRONMENT

The deep features were extracted using Google Colab while the ensemble classification (stacking) was done using Weka 3.9.6. The Anaconda Jupiter Notebook with python programming language was used in these experiments. Moreover, various python libraries including Keras, NumPy, Pandas, and TensorFlow were used in our experiments. The hardware consisted of an NVIDIA Tesla (K80 GPU) with disk storage and graphics memory of 68.40 GB and 16 GB.

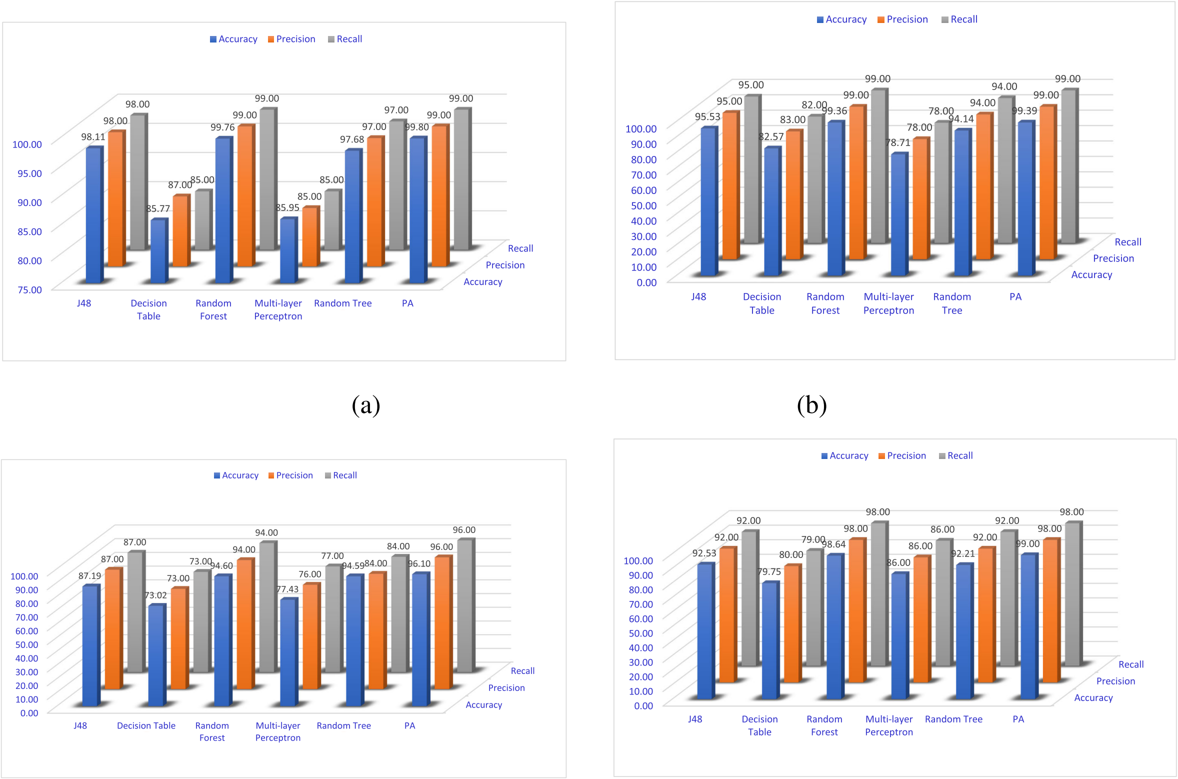
B. ABLATION STUDY

In this ablation study,[1] we measured the performance in terms of percentage accuracy by replacing stacking of PA with various classifiers, namely: decision tree, J48, and random forest [53], [54]. Table 1 shows the ablation study for the PA using SMIC, CASME-II, CAS(ME)^2, and SAMM datasets. It can be seen that there exists a performance degradation when we replace the stacking of PA with any other aforementioned classifiers. The highest degradation occurred when we replaced the stacking with a random forest classifier for the SMIC dataset. Random forest takes much time for training as it combines a lot of decision trees.

C. HYPERPARAMETER TUNING FOR DEEP LEARNING ARCHITECTURES AND CLASSIFIERS

Hyperparameters of deep architectures of the PA are tuned to optimize its performance as shown in Table 2.

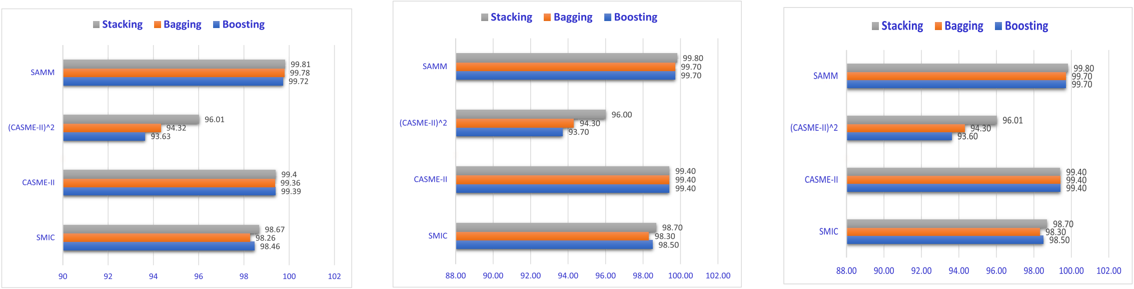
[1] In ablation study, certain parts of the architecture are removed or replaced to study their effect on performance.



(a) (b)

(c) (d)

FIGURE 15. Performance of PA with different classifiers using a) SMIC (3 categories), b) CASME-II (7 categories), c) CAS(ME)^2 (8 categories), d) SAMM (8 categories) datasets.



(a) (b) (c)

FIGURE 16. Stacking of PA is compared with boosting and bagging in terms of (a) accuracy, (b) precision, and (c) recall.

Similarly, the hyperparameters of classifiers of the PA are optimized as can be seen in Table 3.

D. EXPERIMENT 1: ACCURACY COMPARISON WITH OTHER APPROACHES

The comparison of the PA with state-of-the-art approaches, including Tkr [55], XiaB [56], Ond [42], Lng [57], SaiP [36], Zhao and Xu [58], Dvn [59], Yu et al. [60], Dnc [61], Alt [62], and Li et al. [37], in terms of percentage accuracy using SMIC, CASME-II, CASME^2, and SAMM datasets is shown in Fig. 13. It can be observed that PA outperforms the other approaches.

The PA exhibited the best accuracy on SAMM as compared to the other three datasets because it has the highest-resolution images. The highest performance degradation of PA was observed on CASME^2 because this dataset has high intra-class variations and its samples were generated using a relatively low frame rate (30 fps).

Although the images of the SMIC dataset were recorded using FACS coding without neutral sequences and have low resolution, still deep features in combination with ensemble classification (stacking) of PA yielded the best accuracy (98.68%). The second best accuracy was shown by Dnc while Yu showed the worst performance. In CASME-II, the accuracy of PA was 1.13, 1.06, 1.14, 1.15 times higher than

Wg, Adg, Yu, and Dvn while it was 1.05, 1.09, 1.05, 1.41 times better than Tkr, XiaB, Ond, Lng for CASME^2. The accuracy of PA was 1.28, 1.31, 1.22, 1.17 times higher than SaiP, Zhao, Dvn, and Yu using SAMM.

E. EXPERIMENT 2-5: PERFORMANCE COMPARISON OF PA WITH STATE-OF-THE-ART DEEP ARCHITECTURES

As seen in Fig. 14 (a)-(d), our PA is compared with state-of-the-art architectures including DenseNet-121, ResNet-50, and VGG-16 in terms of accuracy, precision, and recall. It can be observed that our PA outperformed the deep architectures in terms of accuracy. The percentage increase of accuracy of PA is 5.43 when compared with ResNet-50 using CASME^2 while this increase is 4.21 as compared to VGG-16 using the SAMM dataset. Both CASME^2 and SAMM are high-resolution datasets as compared to SMIC and CASME-II.

PA exhibited better precision and recall as compared to DenseNet-121, ResNet-50, and VGG-16 using SMIC, CASME-II, CASME^2, and SAMM datasets. The lowest values of precision and recall are shown by VGG-16 and ResNet-50 in the case of the SMIC dataset. The reason might include the low resolution of the SMIC dataset.

F. EXPERIMENT 6-9: PERFORMANCE COMPARISON OF PA USING VARIOUS CLASSIFIERS

As shown in Fig. 15, the ensemble classification (using stacking) of PA was compared with J48, decision table, random forest, multi-layer perceptron, and random tree using four standard datasets, namely: SMIC (three categories), CASME-II (eight categories), CAS(ME)^2 (seven categories), and SAMM (eight categories) [63], [64], [65]. It can be observed that the stacking outperformed the other classifiers in terms of percentage accuracy, precision, and recall. It should be noted that during this comparison, the deep features of the PA remained the same while only the classifiers were compared.

Random forest has shown the second best performance in terms of percentage accuracy, precision, and recall on SMIC, CASME-II, CAS(ME)^2, and SAMM as can be seen in Fig. 15. Random forest is robust to outliers, reduces overfitting, noise, and handles missing values. The multi-layer perceptron i.e., feedforward artificial neural network has shown the worst performance in terms of percentage accuracy, precision, and recall on SMIC, CASME-II, CAS(ME)^2, and SAMM. One of the limitations of multi-layer perceptron is its high training time [66]. It uses a lot of parameters because it is fully connected resulting in redundancy and inefficiency.

G. EXPERIMENT 10-12: PERFORMANCE COMPARISON OF STACKING IN PA WITH BOOSTING AND BAGGING

As shown in Fig. 16, the stacking of PA was compared with boosting and bagging using four datasets, namely: SMIC (three categories), CASME-II (eight categories), CAS(ME)^2 (seven categories), and SAMM (eight categories). It can be observed that stacking outperformed boosting and bagging in terms of percentage accuracy, precision, and recall. It should be noted that the deep features of the PA remained the same, while only the ensemble classification algorithms were compared.

PA showed the highest accuracy of 99.81% on SAMM, which is a high-resolution dataset, as compared to other datasets as shown in Fig. 16 (a). Moreover, the precision and recall of the PA on the SAMM dataset were close to 1 (0.998 and 0.998 respectively) as shown in Fig. 16 (b)-(c). Hence, F1-score was also near 1. Boosting and bagging showed relatively lower accuracy on CAS(ME)^2 as compared to other datasets. One reason includes a larger number of subjects and more micro and macro expressions in this dataset as compared to other datasets.

VI. CONCLUSION AND FUTURE WORK

This study proposed a novel ensemble technique (stacking) in combination with deep features for MER. In order to ensure rigorous experimentation, we compared the PA with 11 existing approaches. It can be concluded that PA outperformed these approaches and achieved an accuracy of 98.68%, 99.39%, 96.01%, and 99.80% on SMIC, CASME-II, CAS(ME)^2, and SAMM datasets respectively. The features of PA were extracted from deep architectures (DenseNet-121, VGG-16, and ResNet-50) and had high dimensions. Therefore, to overcome this problem, we applied PCA for dimensionality reduction. The PA also exhibited better results than the state-of-the-art deep networks including DenseNet-121, ResNet-50, and VGG-16. In order to assess the contribution of stacking in the PA, we compared it with other ensemble techniques and classifiers, namely: boosting, bagging, J48, Decision Table, Random Forest, Random Tree, and feed forward fully connected artificial neural network. The experiments demonstrated that PA outperformed other ensemble techniques and classifiers in terms of percentage accuracy, precision, and recall.

In the future, we would like to extend our work to more challenging environments such as poorly lit areas, crowds, occlusion, and camouflage. The addition of large image repositories will be an important contribution specifically for applications involving deep learning algorithms. Furthermore, we look forward to developing a practical application of the ME detection framework in combination with IoT devices. In the near future, compact deep learning solutions for multiple devices with better accuracy will be in great demand.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Gulnaz Perveen: Implementing Proposed Architecture, Methodology, Software, Gathering Datasets, Experiments and Results, Investigation, Writing, and Editing. **Syed Farooq Ali**: Conceptualization, Designing Deep Architecture, Methodology, Formal Analysis, Discussion on Results, Writing, Resources, Supervision, and Project

administration. **Jameel Ahmad**: Supervision, Writing—original draft, Review, and Gathering Datasets. **Sana Shahab**: Supervision, Review, and Funding. **Muhammad Adnan**: Supervision, Write-Up, Data Gathering, Review, and Proofreading. **Mohd Anjum**: Data Gathering and Funding. **Ikramullah Khosa**: Supervision, Review, Funding, and Proofreading.

REFERENCES

[1] A. A. Varghese, J. P. Cherian, and J. J. Kizhakkethottam, "Overview on emotion recognition system," in *Proc. Int. Conf. Soft-Comput. Netw. Secur. (ICSNS)*, Feb. 2015, pp. 1–5.

[2] A. Chatzimparmpas, R. M. Martins, K. Kucher, and A. Kerren, "StackGenVis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1547–1557, Feb. 2021.

[3] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński, "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal," in *Proc. Asian Conf. Intell. Inf. Database Syst.* Springer, 2010, pp. 340–350.

[4] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

[5] M. Pandey and S. Taruna, "A comparative study of ensemble methods for students' performance modeling," *Int. J. Comput. Appl.*, vol. 103, no. 8, pp. 26–32, Oct. 2014.

[6] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Mach. Learn.*, vol. 54, no. 3, pp. 255–273, Mar. 2004.

[7] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.

[8] X. Ying, "An overview of overfitting and its solutions," *J. Phys., Conf. Ser.*, vol. 1168, Feb. 2019, Art. no. 022022.

[9] L. Wang, Z. He, B. Meng, K. Liu, Q. Dou, and X. Yang, "Two-pathway attention network for real-time facial expression recognition," *J. Real-Time Image Process.*, vol. 18, no. 4, pp. 1173–1182, Aug. 2021.

[10] Y. Feng, X. An, and S. Li, "Research on face recognition based on ensemble learning," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9078–9082.

[11] J. Y. Choi and B. Lee, "Ensemble of deep convolutional neural networks with Gabor face representations for face recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 3270–3281, 2020.

[12] F. Xiong, Y. Xiao, Z. Cao, Y. Wang, J. T. Zhou, and J. Wu, "ECML: An ensemble cascade metric-learning mechanism toward face verification," *IEEE Trans. Cybern.*, vol. 52, no. 3, pp. 1736–1749, Mar. 2022.

[13] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[14] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 6, pp. 636–642, Jun. 1996.

[15] M. Rosenblum, Y. Yacoob, and L. S. Davis, "Human expression recognition from motion using a radial basis function network architecture," *IEEE Trans. Neural Netw.*, vol. 7, no. 5, pp. 1121–1138, Sep. 1996.

[16] Z. Zhang, "Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 13, no. 6, pp. 893–911, Sep. 1999.

[17] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 2, pp. 433–449, Apr. 2006.

[18] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in *Proc. 3rd Int. Conf. Imag. for Crime Detection Prevention (ICDP)*, Dec. 2009, pp. 1–6.

[19] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 28–43, Feb. 2012.

[20] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere, "The Karolinska directed emotional faces: A validation study," *Cognition Emotion*, vol. 22, no. 6, pp. 1094–1118, Sep. 2008.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[22] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1449–1456.

[23] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao, "Micro-expression recognition using color spaces," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6034–6047, Dec. 2015.

[24] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 299–310, Oct. 2016.

[25] Y.-J. Liu, B.-J. Li, and Y.-K. Lai, "Sparse MDMO: Learning a discriminative feature for micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 254–261, Jan. 2021.

[26] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, pp. 1–8, Jan. 2014.

[27] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 254–267, Apr. 2017.

[28] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3160–3172, Nov. 2018.

[29] Q. Li, J. Yu, T. Kurihara, and S. Zhan, "Micro-expression analysis by fusing deep convolutional neural network and optical flow," in *Proc. 5th Int. Conf. Control, Decis. Inf. Technol. (CoDIT)*, Apr. 2018, pp. 265–270.

[30] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "An overview of facial micro-expression analysis: Data, methodology and challenge," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1857–1875, Jul./Sep. 2022.

[31] R. Guermazi, T. Ben Abdallah, and M. Hammami, "Facial micro-expression recognition based on accordion spatio-temporal representation and random forests," *J. Vis. Commun. Image Represent.*, vol. 79, Aug. 2021, Art. no. 103183.

[32] Z. Sun, Z.-P. Hu, M. Zhao, and S. Li, "Multi-scale active patches fusion based on spatiotemporal LBP-TOP for micro-expression recognition," *J. Vis. Commun. Image Represent.*, vol. 71, Aug. 2020, Art. no. 102862.

[33] H. Sadeghi and A.-A. Raie, "Histogram distance metric learning for facial expression recognition," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 152–165, Jul. 2019.

[34] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108275.

[35] Z. Zhang, T. Chen, H. Meng, G. Liu, and X. Fu, "SMEConvNet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos," *IEEE Access*, vol. 6, pp. 71143–71151, 2018.

[36] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[37] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 563–577, Oct. 2018.

[38] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 626–640, Mar. 2020.

[39] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184537–184551, 2019.

[40] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.

[41] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 424–436, Oct. 2018.

[42] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018.

[43] I. Kareem, S. F. Ali, and A. Sheharyar, "Using skeleton based optimized residual neural network architecture of deep learning for human fall detection," in *Proc. IEEE 23rd Int. Multitopic Conf. (INMIC)*, Nov. 2020, pp. 1–5.

[44] M. K. Yaqoob, S. F. Ali, I. Kareem, and M. M. Fraz, "Feature-based optimized deep residual network architecture for diabetic retinopathy detection," in *Proc. IEEE 23rd Int. Multitopic Conf. (INMIC)*, Nov. 2020, pp. 1–6.

[45] T. Abbas, S. F. Ali, A. Z. Khan, and I. Kareem, "OptNet-50: An optimized residual neural network architecture of deep learning for driver's distraction," in *Proc. IEEE 23rd Int. Multitopic Conf. (INMIC)*, Nov. 2020, pp. 1–5.

[46] T. Abbas, S. F. Ali, M. A. Mohammed, A. Z. Khan, M. J. Awan, A. Majumdar, and O. Thinnukool, "Deep learning approach based on residual neural network and SVM classifier for driver's distraction detection," *Appl. Sci.*, vol. 12, no. 13, p. 6626, Jun. 2022.

[47] M. K. Yaqoob, S. F. Ali, M. Bilal, M. S. Hanif, and U. M. Al-Saggaf, "ResNet based deep features and random forest classifier for diabetic retinopathy detection," *Sensors*, vol. 21, no. 11, p. 3883, Jun. 2021.

[48] X. Chen, X. Lou, L. Bai, and J. Han, "Residual pyramid learning for single-shot semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 2990–3000, Jul. 2020.

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

[50] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[51] G. Huang, Z. Liu, G. Pleiss, L. van der Maaten, and K. Q. Weinberger, "Convolutional networks with dense connectivity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8704–8716, Dec. 2022.

[52] H. Mustafa, S. F. Ali, M. Bilal, and M. S. Hanif, "Multi-stream deep neural network for diabetic retinopathy severity classification under a boosting framework," *IEEE Access*, vol. 10, pp. 113172–113183, 2022.

[53] J. Liu, H. Wang, and Y. Feng, "An end-to-end deep model with discriminative facial features for facial expression recognition," *IEEE Access*, vol. 9, pp. 12158–12166, 2021.

[54] S. Sheikholeslami, M. Meister, T. Wang, A. H. Payberah, V. Vlassov, and J. Dowling, "AutoAblation: Automated parallel ablation studies for deep learning," in *Proc. 1st Workshop Mach. Learn. Syst.*, Apr. 2021, pp. 55–61.

[55] M. A. Takalkar, M. Xu, and Z. Chaczko, "Manifold feature integration for micro-expression recognition," *Multimedia Syst.*, vol. 26, no. 5, pp. 535–551, Oct. 2020.

[56] B. Xia, W. Wang, S. Wang, and E. Chen, "Learning from macro-expression: A micro-expression recognition framework," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2936–2944.

[57] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[58] Y. Zhao and J. Xu, "Compound micro-expression recognition system," in *Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS)*, Jan. 2020, pp. 728–733.

[59] A. Davison, W. Merghani, and M. Yap, "Objective classes for micro-facial expression recognition," *J. Imag.*, vol. 4, no. 10, p. 119, Oct. 2018.

[60] J. Yu, C. Zhang, Y. Song, and W. Cai, "ICE-GAN: Identity-aware and capsule-enhanced GAN with graph-based reasoning for micro-expression recognition and synthesis," 2020, *arXiv:2005.04370*.

[61] R. Danescu, D. Borza, and R. Itu, "Detecting micro-expressions in real time using high-speed video sequences," in *Intelligent Video Surveillance*. London, U.K.: IntechOpen, 2018.

[62] B. Allaert, I. M. Bilasco, and C. Djeraba, "Micro and macro facial expression recognition using advanced local motion patterns," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 147–158, Jan. 2022.

[63] S. Ali, R. Khan, A. Mahmood, M. Hassan, and A. Jeon, "Using temporal covariance of motion and geometric features via boosting for human fall detection," *Sensors*, vol. 18, no. 6, p. 1918, Jun. 2018.

[64] S. F. Ali, A. S. Aslam, M. J. Awan, A. Yasin, and R. Damaševičius, "Pose estimation of driver's head panning based on interpolation and motion vectors under a boosting framework," *Appl. Sci.*, vol. 11, no. 24, p. 11600, Dec. 2021.

[65] S. Aftab, S. F. Ali, A. Mahmood, and U. Suleman, "A boosting framework for human posture recognition using spatio-temporal features along with radon transform," *Multimedia Tools Appl.*, vol. 81, pp. 42325–42351, Aug. 2022.

[66] H. Zheng, G. Wang, and X. Li, "Swin-MLP: A strawberry appearance quality identification method by Swin transformer and multi-layer perceptron," *J. Food Meas. Characterization*, vol. 16, pp. 2789–2800, Apr. 2022.

GULNAZ PERVEEN received the M.Sc. degree in computer science from Bahauddin Zakariya University (BZU), Multan, Pakistan. She is currently pursuing the M.Sc. degree in computer science with the School of Systems and Technology, University of Management and Technology, Lahore, Pakistan. She is currently a freelance software engineer. Her current research interests include machine learning, computer vision, deep learning, and pattern recognition.

SYED FAROOQ ALI received the M.S. degree (Hons.) in computer science from LUMS, Lahore, Pakistan, the M.S. and Ph.D. degrees in computer science from Ohio State University, Columbus, USA, and the Ph.D. degree in computer science from UMT, Pakistan. He held a LUMS Fellowship. He is currently an Assistant Professor with UMT. His current research interests include computer vision, digital image processing, and medical imaging. He is the reviewer of various IEEE conferences and journals.

JAMEEL AHMAD received the M.Sc. degree in electrical engineering from the University of Southern California at Los Angeles, USA, and the Ph.D. degree in electrical engineering from the UET Lahore. He was with Qualcomm and Broadcom, San Diego, CA, USA, where he is focusing on 3G mobile communication systems. Currently, he is an Assistant Professor with the University of Management and Technology, Lahore, Pakistan. His current research interests include parallel and distributed computing, machine learning, and optimization of smart microgrids for energy internet. He is the reviewer of numerous journals and conferences.

SANA SHAHAB is currently an Assistant Professor with the College of Business Administration, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her current research interests include the interdisciplinary applications of statistics and management science to serve the broad areas of problem-solving and decision-making in the organization, machine learning, deep learning, and the Internet of Things. She has published and presented more than 40 research articles in reputed journals and international conferences in her research area.

MUHAMMAD ADNAN received the bachelor's degree in electrical engineering (electronics and communication) from UCET, Bahuaddin Zakariya University, Multan, in 2002, the M.S. degree in telecommunication engineering from UET, Peshawar, in 2005, and the Ph.D. degree in information and communication systems from the Beijing University of Posts and Telecommunication, China, in 2010. He was with the PTCL-Etisalat Academy, from 2004 to 2010, and the NFC Institute of Engineering and Technology, Multan, from 2011 to 2013. His current research interests include information and communication, energy, and artificial intelligence. He is the reviewer of numerous journals and conferences.

MOHD ANJUM received the M.Tech. degree in computer science and engineering (software engineering) and the Ph.D. degree in computer engineering from Aligarh Muslim University, India. He was an Assistant Professor with Aligarh Muslim University, from 2012 to 2015. His current research interests include management, the Internet of Things, and machine learning. He has published and presented numerous research papers in reputed journals and international conferences in his area of interest.

IKRAMULLAH KHOSA (Member, IEEE) received the Ph.D. degree in electronics and telecommunications from Politecnico di Torino, Italy, in 2015. He is currently an Assistant Professor with the Electrical and Computer Engineering Department, COMSATS University Islamabad, Lahore Campus. His current research interests include artificial intelligence, data analysis, machine learning, and pattern recognition.