IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 19 May 2023, accepted 11 June 2023, date of publication 19 June 2023, date of current version 3 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287389

RESEARCH ARTICLE

Generalizing a Small Facial Image Dataset Using Facial Generative Adversarial Networks for Stroke's Facial Weakness Screening

PHONGPHAN PHIENPHANICH[1,2], (Student Member, IEEE), NICHAPA LERTHIRUNVIBUL[2], EKABHAT CHARNNARONG[2,3], ADIREK MUNTHULI[1,2], CHARTURONG TANTIBUNDHIT[1,2], (Member, IEEE), AND NIJASRI C. SUWANWELA[4]

[1]Department of Electrical and Computer Engineering, Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Rangsit Campus, Khlong Luang, Khlong Nueng, Pathum Thani 12120, Thailand
[2]Center of Excellence in Intelligent Informatics, Speech and Language Technology, and Service Innovation (CILS), Thammasat University, Rangsit Campus, Khlong Luang, Khlong Nueng, Pathum Thani 12120, Thailand
[3]Patumwan Demonstration School, Pathum Wan, Bangkok 10330, Thailand
[4]Department of Medicine, Faculty of Medicine, Chulalongkorn University, Pathum Wan, Bangkok 10330, Thailand

Corresponding author: Charturong Tantibundhit (tchartur@engr.tu.ac.th)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Chulalongkorn University under Application No. 242/61.

ABSTRACT Stroke is a medical emergency resulting from disruption of blood supply to different parts of the brain which leads to facial weakness and paralysis as the brain is the control center. Stroke is the leading cause of long-term disability which significantly changes the patient's life. This paper introduces the use of facial image dataset containing neutral and smiling expressions to classify facial weakness which is a common sign of stroke. Our ''real facial image dataset'' comprises of face images of normal subjects and stroke patients. However, to increase the dataset, we added another dataset known as ''FaceGAN dataset''. This additional dataset contains a pair of neutral and smiling facial image synthesized from public datasets which were augmented to generate two additional smiling images at eight different age groups. The faces were divided into left and right side using facial landmark detection technique and corrected for geometric distortions through affine transformation matrix from Delaunay triangulation. An autoencoder model composed of ConvNeXt encoder and ConvNet decoder was trained and used to fine-tune a facial weakness classification model from our proposed architecture. Results from four-fold cross validation showed that the model validation was less prone to overfitting when used with the FaceGAN dataset, with an average AUC of 0.76 and F1-score of 71.19%, compared to without FaceGAN data which only achieved an F1-score of 61.54%. This study shows that the FaceGAN can efficiently generalize models for programs with a small dataset for use with stroke detection. This work can be further improved and optimized for clinical application in the future.

INDEX TERMS Facial generative adversarial networks, facial weakness, FAST, small dataset, stroke-screening.

I. INTRODUCTION

Stroke is a medical emergency that requires immediate attention as it can become a long-term disability. Moreover, it is the second leading cause of death worldwide [1]. Screening for stroke includes history taking, physical examination and assessment of risk factors such as age or certain cardiovascular diseases. Signs and symptoms that are indicative of stroke include facial weakness, arm or leg weakness or numbness, and slurred speech [2], [3], [4]. Hence, the Face, Arm, Speech, Time (FAST) assessment was devised to stress the importance of timely diagnosis and management of stroke. Timely recognition of a stroke can halt the rate of neurons loss from the process of ischemia as emphasized by the phrase ''time is brain'' [5].

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

Early diagnosis and prompt treatment of stroke can reduce disease morbidity and mortality. Hence, fast recognition of stroke is crucial in reducing subsequent post-stroke disabilities [1]. However, diagnosis may be deferred due to several factors such as delayed recognition or lack of access to appropriate medical equipment required for diagnosis. Incorporation of screening tools using digital images may provide faster assessment and detection of stroke enabling improved post-stroke outcomes. In our previous work [3], [6], we created a self-screening tool for stroke using the gyroscope and accelerometer functions in smartphones to detect arm weakness.

Other than weakness of extremities, facial palsy is also a sign of stroke. In some cases, weakness may be accompanied by numbness or difficulty speaking [1]. Facial muscle weakness can be presented as asymmetry of facial features or differences in the size or shape of one side of the face compared to the other. Signs of facial palsy include drooping of the eyelid, corner of the mouth, or lower lip. Patients may also have difficulty moving certain facial muscles which may be more pronounced when they are asked to smile or raise their eyebrows. However, facial weakness may be subtle making it difficult to detect especially for non-neurologists [7]. Moreover, Brandler et al. [8] demonstrated that paramedics failed to identify facial weakness in 17% of stroke patients and incorrectly diagnosed facial weakness in 33% of cases.

A common differential diagnosis of acute stroke that presents with unilateral facial weakness is Bell's palsy. Unlike stroke which affects the upper motor neuron of the facial nerve, Bell's palsy is a lower motor neuron facial palsy. Clinical differences between the two conditions include involvement of upper facial muscle paralysis that can be found in Bell's palsy but is usually absent in stroke. Clinical examination is warranted to localize the lesion. Weakness of the lower face sparing the forehead suggests an upper motor neuron lesion, whereas weakness of both the lower and upper face indicates a lower motor neuron lesion [9].

Other differences include onset of symptoms and associating signs and symptoms such as arm or leg weakness in stroke or decreased lacrimation and salivation in Bell's palsy. Treatment of acute stroke is initiation of thrombolysis. On the other hand, Bell's palsy is treated with early administration of steroids and eye protection to prevent ocular complications. Hence, it is crucial to correctly differentiate between the two etiologies of facial palsy to avoid misdiagnosis and delayed appropriate treatment, and also to prevent overreliance on neuroimaging [9].

Due to the unprecedented COVID-19 outbreak, all subjects had to wear masks at all times in the hospital and we had to discontinue our data collection process decreasing the number of facial images. Furthermore, concerns regarding subject's privacy and confidentiality, as our work focused on facial features, also limited the amount of data collected. Hence, we had to find other alternatives to increase and generalize our dataset. Other options include adding images and videos from public data sources like Google or Youtube [4], which we did not incorporate because these real-world images come in different image sizes and resolution making it difficult to use with our program. Instead, we chose to use Facial Generative Adversarial Networks (FGAN) which is a deep-learning and auto-encoder model trained from a large-scale collection of face images to generate new face images [10].

FGAN combines specific features and random values, called latent vector, to create new face images that closely resembles the actual image which poses a challenge for the discriminant model to classify between real and fake face images. These models [10], [11] can generate a wide variety of facial appearances such as age, skin-color, ethnicity, emotional expressions, and facial features such as facial hair and glasses. Therefore, we chose FGAN over public databases to increase our dataset and generalize our model. Furthermore, FGANs have been used to reconstruct 3D heads in full 360-degree view from 2D facial images as demonstrated by PanoHead [12].

We recognize the importance of a timely diagnosis of stroke for improved patient outcomes. Hence, we propose an algorithm to create a screening tool for early stroke detection with good performance using two facial images, a neutral and smiling state. Moreover, we aim to develop a tool that is easily accessible by smartphones and tablets without requiring much resources. During the process, we also incorporated FGANs to synthesize more facial images and alternative smiling phases which overcomes the limitations in datasets.

II. RELATED WORKS

Facial weakness detection from a single image can be done using a simple programming model which involves three main processes. Firstly, the model identifies the person's face and mouth. Facial landmark detection plays an important role in locating the face then rotating it to a neutral position. A widely-used facial landmark model, the 68-landmarks model [2], [4], [13], [14], [15], [16], [17] allocates 20 points in outlining the shape of the mouth that can be used to detect lip drooping. Other work [18] uses the 106-landmarks model which differs from the 68-landmarks model in that it assigns more points to the eyes. However, both models have the same efficacy in detecting changes in lip shape.

Feature extraction is the next step. Many programs use hand-craft features which analyze the distances between landmark points such as the relation between the corner of the mouth to the eye [13], [16], [17], [18]. This technique is efficient but heavily depends on accurate detection of facial landmarks. Therefore, some works [19], [20] required manual re-annotation on images of facial palsy patients because the facial landmarks models were trained on healthy subjects in public datasets. Many studies [2], [4], [17] employ the histogram of oriented gradients (HOG) for facial recognition.

HOG is a technique that counts occurrences of gradient orientation in a selected portion of an image, then uses the principal component analysis (PCA) method to simplify the image into a lower dimensional space. Other works [15], [21] use deep learning techniques such as Convolutional Neural Network (CNN), to embed face region to features vector.

The HOG and deep learning techniques require training on an extensive amount of data points in order to create an embedded space that can adequately differentiate between normal and facial palsy subjects. A study by Zhuang et al. [4] included 437 subjects with diverse demographics and separated subjects based on skin color and gender. However, many studies [13], [14], [15], [16], [17] including this study recruited less than 200 subjects from a more specific demographic. However, to overcome the limited dataset, data augmentation was employed. Data augmentation [22] enables iterative use of data points by flipping the facial image and adjusting image brightness and contrast among other methods.

The last step is using features vector to classify facial weakness. Most studies [2], [15], [16], [17], [18], [19] employ machine learning techniques, such as support vector machine (SVM), random forest, and k-nearest neighbors (kNN), to classify facial abnormalities. These machine learning programs analyze one single image making the process compact and fast. However, there are limitations in accurately detecting facial landmarks and selecting the target image. Additionally, healthy subjects in real-world samples usually have asymmetrical faces in neutral position which can be further challenged by various facial expressions. A feature-based recognition model designed for expressive facial images had lower accuracy in identifying the nose and mouth areas as compared to the eyes in expressive image variants [23].

As opposed to using static images, training models with dynamic facial expression videos can enhance its performance at the expense of reduced reliability [24]. Facial weakness classification from video clips requires a more complex process which involves localization of facial features and movements in each frame from neutral to smiling state to increase accuracy. Zhuang et al. [4], improved their accuracy by assigning frame features to recurrent neural network (RNN), such as bidirectional long short-term memory (BiLSTM). Classification from videos require more resources making it incompatible with our previous work [3] which aims for use with budget smartphones to increase accessibility.

Deng et al. [25] proposed a model with improved robustness and accuracy of facial recognition from large-scale databases by using angular margin loss which represents the differences in angles of embedded vector with margin penalty from the facial image. Their work inspired us to take advantage of the latent vector of auto-encoder model for facial weakness detection from different angles of vectors embedded from each side of the face in neutral and smiling images. Furthermore, we employ the auto-encoder model, pre-trained and augmented using the FGAN dataset, to embed near-mouth region from each side of the face. Then, we use our proposed comparison vector layer and artificial neural network (ANN) to classify facial weakness.

Some studies created algorithms for facial weakness detection using video analysis which differs from this study [4], [26]. In this paper, we aim to enhance the facial recognition model with the advantage of video classification which preserves the dynamic changes from neutral to smiling state while maintaining the simplicity of using one single image. We propose the use of two-state images which are the neutral and smiling faces. Our proposed algorithm cannot be compared with other video classifications as it requires only two images making it practical and convenient. It is crucial we create a simple algorithm as we aim to create an application that can easily be accessed by tablets and smartphones.

The rest of the paper includes data collection, generation and pre-processing. Data collection outlines the types of participants we recruited and the challenges in working with participants due to the COVID-19 outbreak. The resolution to this unforeseen issue is explained in latter sections. Followed by the proposed method in developing a facial classification model, then the experimental setup and results. Lastly, the conclusion and discussion along with acknowledgements.

III. DATA COLLECTION, GENERATION, AND PRE-PROCESSING

A. DATA COLLECTION

This study was designed to collect two frontal facial images from 63 participants including 12 normal control subjects, 27 chronic stroke patients without facial weakness, and 24 stroke patients with facial weakness ranging from slight to pronounced weakness. The data collection was approved by the Institutional Review Board of the Faculty of Medicine at Chulalongkorn University with approval number 242/61, and all participants provided informed consent prior to data collection at King Chulalongkorn Memorial Hospital, Bangkok, Thailand. The images were captured using a smartphone device camera with a resolution of 1,920 × 1.080 pixels with neutral facial expression images and smiling facial expression images. The term ''real subjects dataset'' is used throughout the paper to describe this dataset.

B. GENERATING DATA WITH FACIAL GENERATIVE ADVERSARIAL NETWORK

Generative Adversarial Networks (GAN) is a variant of unsupervised deep neural networks, most commonly used for generating artificial data with the same statistical properties as the training data [27]. GAN consist of two different networks, referred to as a generator and a discriminator, which are used to generate images and determine whether the generated images are realistic or not, respectively. These

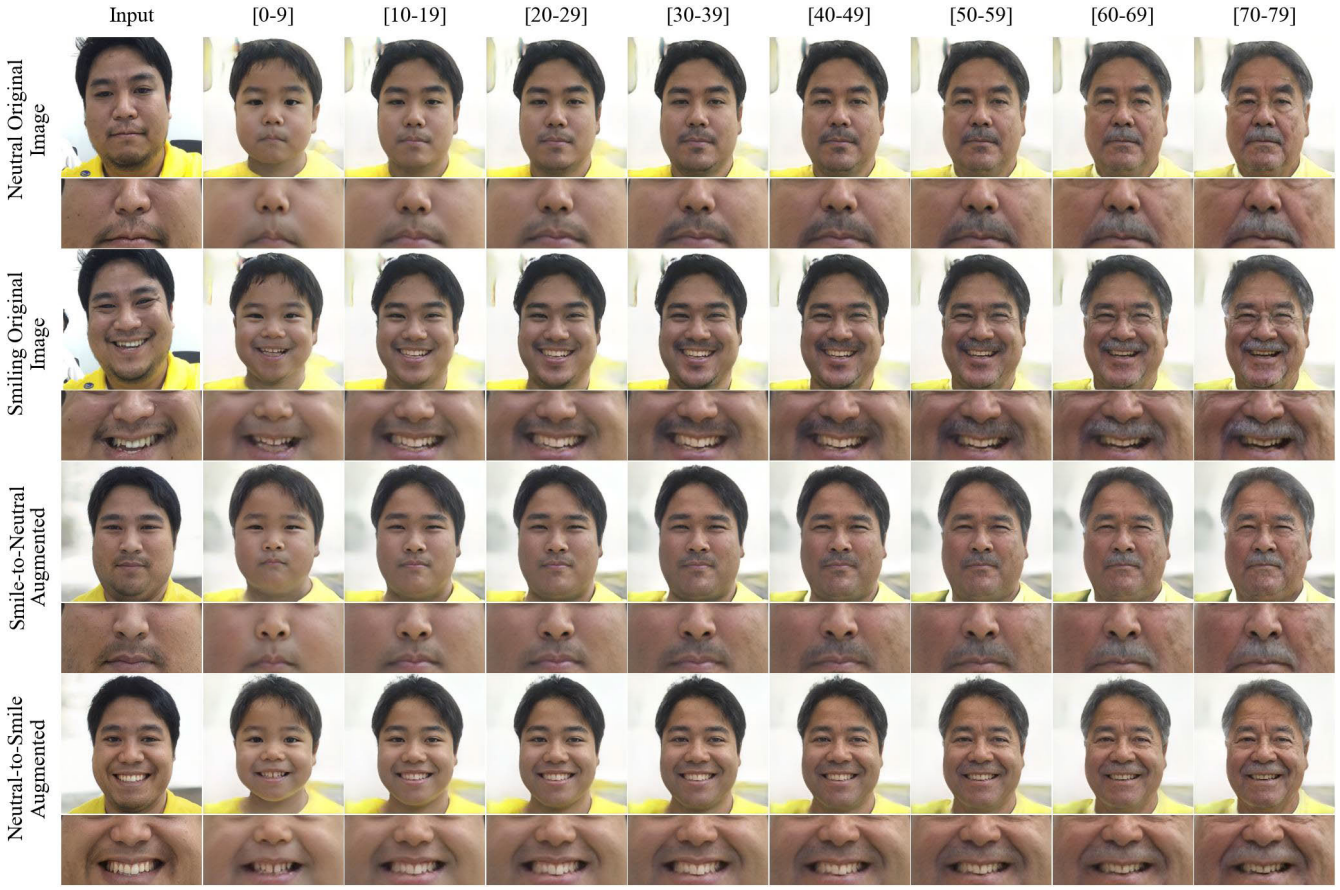

FIGURE 1. Using StyleCLIP GAN [10] and OverLORD GAN [11] to generate two additional smiling phases across eight distinct age ranges from an original facial image.

two models are trained simultaneously with the generator attempting to generate data that is indistinguishable from real data, and the discriminator attempting to differentiate between real and generated data [27].

FGAN is a subset of GAN that is can generate realistic facial images from a given training dataset [28]. This generative model can be used to create a variety of facial images with different ages, genders, and ethnicities [10], [11]. Furthermore, FGAN can be used to remove facial defects such as blemishes and wrinkles, as well as generate images of fictional characters that are widely used in many social media platforms [10], [11].

This paper utilizes 1,600 synthetic neutral facial expression images from the Generated Photos dataset [29], inspired by concepts of GAN and StyleGAN. The images consist of 400 faces of different age groups (child, young-adult, adult, and elderly) with varying skin color, ethnicity, and appearance. These images were augmented using a pre-trained StyleCLIP GAN [10] to generate an additional two smiling (smile-to-neutral and neutral-to-smile augmentation) images and a pre-trained OverLORD GAN [11] to generate eight additional age ranges, apart from the original image. This resulted in a total of 14,400 pairs of neutral and smiling facial expression images, as illustrated in Fig. 1. The term "FaceGAN" is used throughout the paper to represent this dataset.

C. FACIAL LANDMARK DETECTION AND AFFINE TRANSFORM

Facial landmark detection is a computer vision technique which can identify and locate points corresponding to specific facial features. This technique has been implemented in a vast array of applications including facial recognition, facial classification, 3D face modeling, and facial transformation [30], [31]. A state-of-the-art 2D/3D facial landmark technique is RetinaFace [32], which is an open-source facial recognition system. RetinaFace [32] employs a multi-task learning deep convolutional neural network to detect and locate five important facial landmarks, including the eyes, nose, and mouth. This system has the capability to detect faces even in challenging conditions, such as varied lighting, poses, and facial expressions [32].

Affine transformation is a combination of linear transformations such as rotation, reflection, scaling, shearing, and translation, which are widely used to subtly modify an original image while preserving its overall structure [33]. Facial landmark detection combined with affine transformation can be used to reproduce realistic facial images by making minor

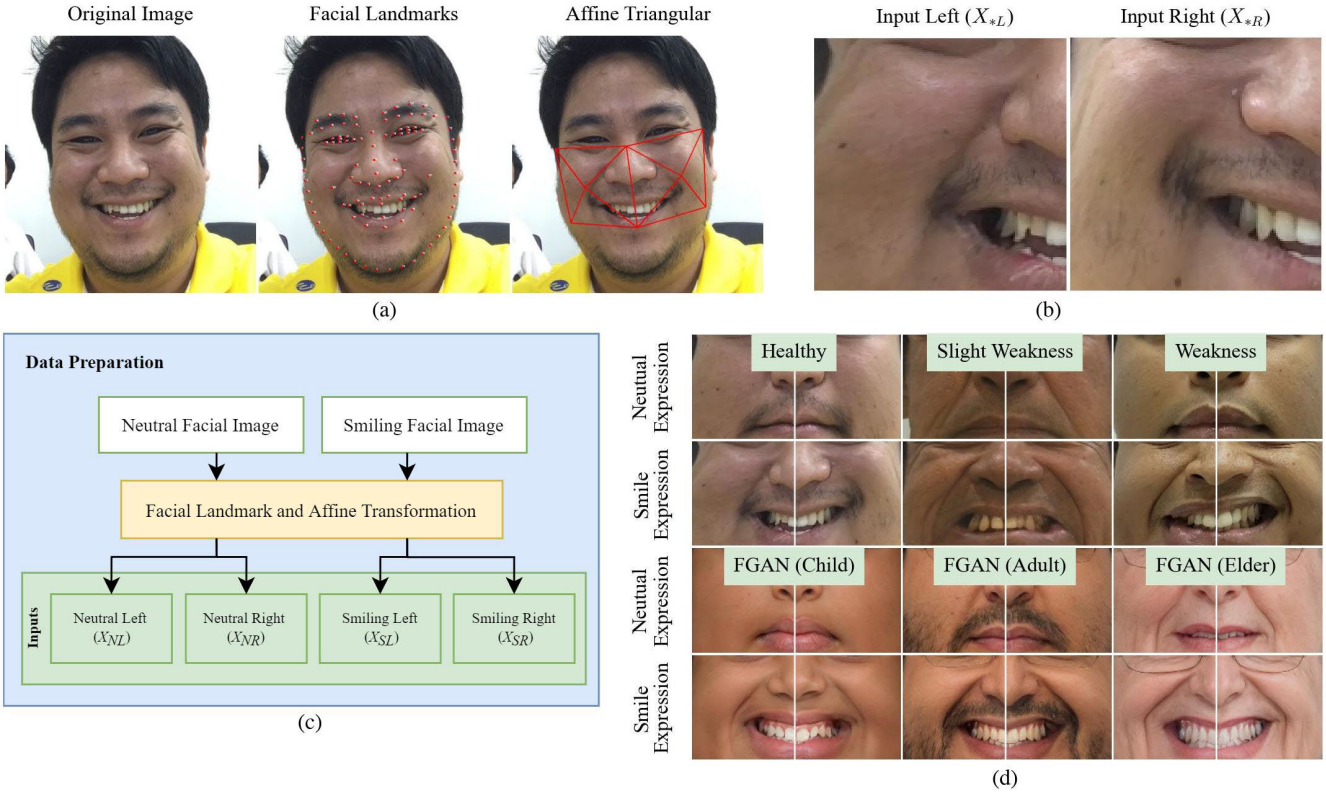

FIGURE 2. The data pre-processing process includes (a) the original image, with facial landmarks from RetinaFace, and with eight triangular matrices from Delaunay triangulation, (b) side-by-side left and right-hand sides of facial images, (c) data augmentation using facial landmark and affine transformation on a pair of neutral and smiling images to generate FaceGAN dataset and (d) six FaceGAN images created with affine transformation matrices from a pair of neutral and smiling facial images.

adjustments such as facial types and alignment [34]. Furthermore, facial landmark detection and affine transformation are also used for facial recognition and matching of one's face to another image [35].

This work used a RetinaFace model [32] to extract 106 facial landmarks. Six out of the 106 points were selected corresponding to the outer corners of the face, center of the lower lip, and center of the nose, as shown in Fig. 2 a). The six facial landmarks were divided into two halves using a vertical line between the two center points. A centroid was created for each of the four points on each side. Five points from each side were used to create a mesh of triangles from a set of points using Delaunay triangulation techniques [36], as in Fig. 2 a), to generate eight affine transformation matrices. Each half side of the image was then applied with the affine transformation matrix to recreate side-by-side left and right-hand sides of facial images, as in Fig. 2 b).

The proposed technique was implemented on FaceGAN dataset to create an input for this work. Since FaceGAN consists of a pair of neutral and smiling facial expression images, the model's input is a 4-dimensional vector $X_{\mathrm{T}} = \{X_{\mathrm{NL}}, X_{\mathrm{NR}}, X_{\mathrm{SL}}, X_{\mathrm{SR}}\}$ composed of the four different images, i.e., the left-hand sides of the neutral facial expression images X_{NL}, the right-hand sides of the neutral facial expression images X_{NR}, the left-hand sides of the smiling facial expression images X_{SL}, and the right-hand sides of the smiling facial expression images X_{SR}, as illustrated in Fig. 2 c)−Fig. 2 d).

IV. PROPOSED METHOD

A. PRE-TRAINING OF AUTOENCODER-BASED UNSUPERVISED MULTI-TASK LEARNING FOR NEUTRAL AND SMILING FACIAL EXPRESSION IMAGES

In this study, ConvNeXt [37] layers were used as the primary layers of the encoder ($\mathcal{E}(\cdot)$) of an autoencoder. The standard CNN layers (ConvNet) were employed as the decoder ($\mathcal{D}(\cdot)$) to map the latent vector of the encoded representation back to the original input space. ConvNeXt architecture enables the latent vector space to take advantage of its strong spatial information processing capabilities, while ConvNet is effective at reconstructing the input data. The results of our experiment demonstrate that this particular combination is most suitable for our work when compared with other architectures, such as the ResNet architecture, or the use of both ConvNeXt as both encoder and decoder for the autoencoder.

For the encoder, we employed a four-stage ConvNeXt design. Each of the four stages consists of 3, 3, 6, and 3 ConvNeXt blocks with 16, 32, 64, and 128 filters, respectively. To reduce overfitting, we incorporated stochastic depth [38] with probabilities of 0.75, 0.5, 0.5, and 0.5 for each

ConvNext block in the four stages, respectively. After the final ConvNeXt block, we included an adaptive average pooling layer that flattens the output to generate a feature vector. This feature vector was then processed through a fully connected layer creating a 48-dimensional latent vector. For the encoder model's initialization, we utilized the Glorot normal initializer [39].

The decoder was created using the latent vector obtained from the encoder with a fully connected layer to project the input. The input was modified to match the output architecture of the adaptive average pooling layer. Then, we implemented a sequence of six deconvolutional (transposed convolutional) layers with filter counts of 64, 64, 32, 32, 24, and 24, to achieve an upsampling of the original input dimensions of the encoder. Additionally, we added convolutional layers with an equivalent number of filters before exponential linear unit (ELU) [40] activation functions after each deconvolutional layer.

Figure 3 a) illustrates the use of facial images from the data loader during the pre-training phase to learn the encoder and decoder representations of the input. This phase requires an input of the entire image of a neutral (X_N) or smiling face (X_S). Two loss functions are calculated: the reconstruction loss ($\mathcal{L}_{RE}$) and the angular distance loss ($\mathcal{L}_{ANG}$). $\mathcal{L}_{RE}$ is the mean average error between the vector of the original input image from FaceGAN and the reconstructed input obtained from the decoder component of an autoencoder, as expressed in Eq. 1. The angular distance loss measures the modified arccos similarity between the latent vector randomizer derived from the metadata of the input and the latent vectors generated by the encoder of an autoencoder from the whole facial image input, as stated in Eq. 3.

$$\mathcal{L}_{RE} = \|X - \mathcal{D}(\mathcal{E}(X))\|_2 \tag{1}$$

$$\text{sim}(V_1, V_2) = \frac{V_1}{\|V_1\|_2} \cdot \frac{V_2}{\|V_2\|_2} \tag{2}$$

$$\mathcal{L}_{ANG} = \frac{1}{\pi} \cdot \arccos(\text{sim}(V_t, \mathcal{E}(X))) \tag{3}$$

$$\mathcal{L}_T = \alpha_{RE} \cdot \mathcal{L}_{RE} + \alpha_{ANG} \cdot \mathcal{L}_{ANG}, \tag{4}$$

where α_{RE} is 1 and α_{ANG} is 10.

We employ a randomized non-negative matrix factorization approach to obtain a set of orthogonal basis vectors and select the nearest orthogonal vector for each latent vector attribute. Cosine similarity is then used to compare the degree of smiling and limit the cosine similarity to positive values. To avoid overfitting, a small amount of Gaussian noise is added to the latent vector. Finally, an Adam optimizer with 0.001 learning rate is used to optimize a total loss ($\mathcal{L}_T$) consisting of the reconstruction loss ($\mathcal{L}_{RE}$) and angular distance loss ($\mathcal{L}_{ANG}$) [41], as shown in Eq. 4. The loss function is used to minimize the distance between the latent vector representation of a FaceGAN image and its metadata. This enables differentiation of trajectory between a normal facial expression and any degree of smiling facial expressions, as well as the difference in age in images that were augmented by FaceGAN.

B. MULTI-TASK LEARNING OF NEUTRAL AND SMILING FACIAL EXPRESSIONS FOR DETECTION OF FACIAL WEAKNESS

In this subsection, a 4-dimensional vector $X_T = \{X_{NL}, X_{NR}, X_{SL}, X_{SR}\}$ representing the input facial image is forwarded to the encoder layer of the autoencoder model derived from the pre-training phase to compute a latent vector representation for each facial image, i.e., a set of vectors, $V_T = \{V_{NL}, V_{NR}, V_{SL}, V_{SR}\}$. These latent vectors V_T are then forwarded to the vector comparison state layer to compute the state vector which is composed of five similarity vectors.

The cosine similarity layer employs a modified cosine similarity function with trainable parameters θ ($\mathcal{C}$), which is defined by a Gaussian function ($\mathcal{G}$) with α scaling and θ shift, as shown in Eq. 5–6. Vector comparison state layer utilizes this function to compare latent vectors called state vectors. For pairs of vectors that should be similar, the θ in the cosine similarity layer should be close to zero, such as V_{NL} and V_{NR}, V_{SL} and V_{SR}, and $V_{SL} - V_{NL}$ and $V_{SR} - V_{NR}$. For pairs of vectors that should be different, the θ in the layer should be close to $\pi/2$, as illustrated in Fig. 4.

$$\mathcal{G}(x;\ \alpha, \theta) = \exp(-0.5 \cdot (\frac{x - \theta}{\alpha})^2) \tag{5}$$

$$\mathcal{C}(V_1, V_2;\ \theta) = \mathcal{G}(\arccos(\text{sim}(V_1, V_2));\ \alpha = \pi/4, \theta) \tag{6}$$

In order to minimize the difference between the baseline output of the ConvNeXt encoder and the output of the state vector layer, a state vector loss ($\mathcal{L}_{SV}$) is computed by calculating Hinge loss between the state vector from the vector comparison state layer and a ground truth state vector, which is based on the similarity of each state vector (similar degree = 0; different degree = 1). Additionally, a state vector is passed onto an ANN to construct a facial weakness classification model and compute a facial weakness loss ($\mathcal{L}_{FL}$) between the ground truth label and the classification result.

The latent vectors V_T are forwarded to the ConvNet decoder layer of the pre-trained autoencoder model in order to reconstruct the original image and calculate the reconstruction loss ($\mathcal{L}_{RE}$) between the input image from the data loader and the reconstructed image. During the multi-task learning training phase, all three losses are computed simultaneously to reduce model sum of weight loss, with $\alpha_{RE} = 0.1$, $\alpha_{FL} = 1$, and $\alpha_{SV} = 5$ while incorporating Adam optimizer to minimize total loss with a 0.0005 learning rate. This study builds two variants of the model: one containing only the real subjects dataset, and one containing both the real subjects and the FaceGAN datasets. We employ the TensorFlow 2.10 framework for model development, utilizing a batch size of 24 during the training process. Our computational resources consist of a 16-core CPU, 64 GB of RAM, and an NVIDIA RTX 3060 GPU with 12 GB of RAM.

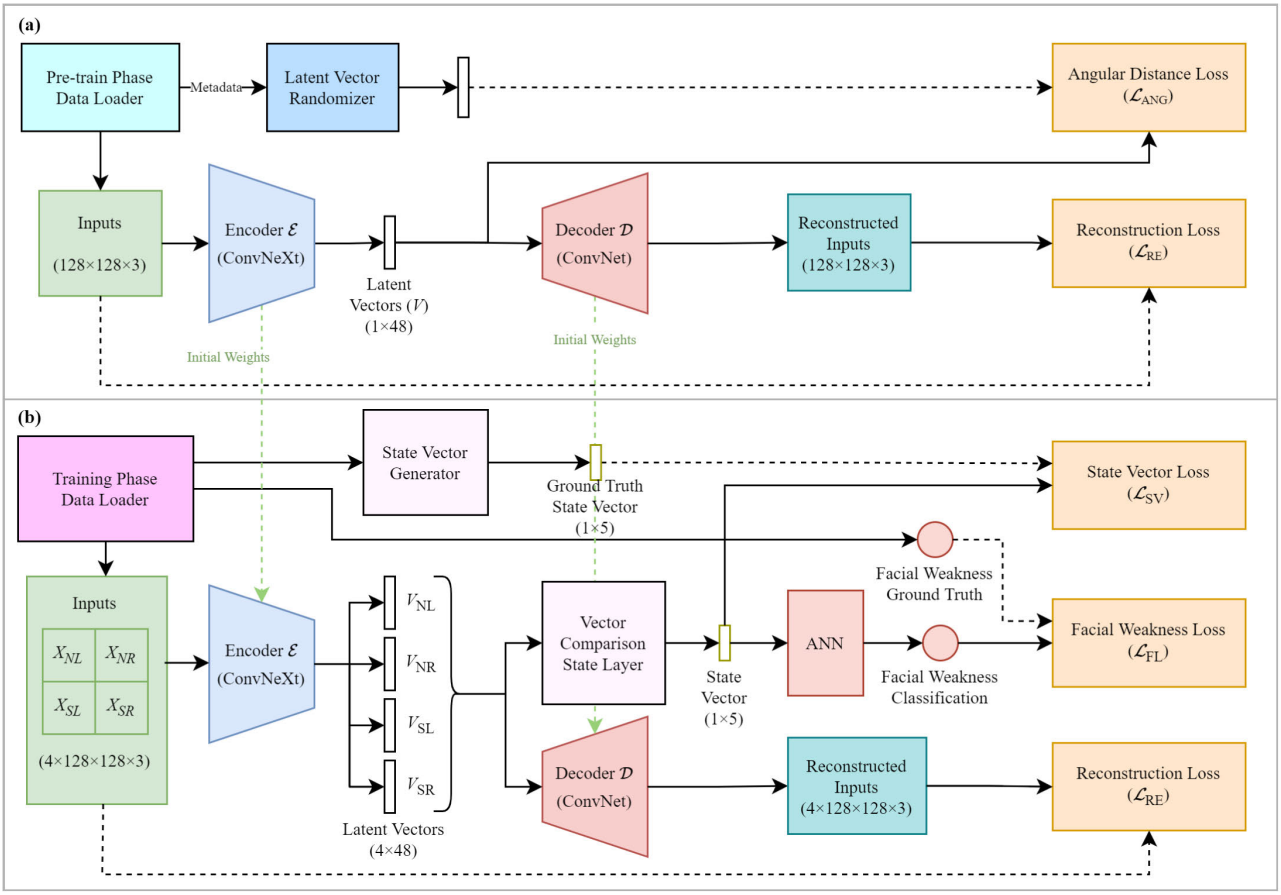

FIGURE 3. The proposed model consists of two parts: a) a pre-training phase for learning the encoder and decoder representation of the FaceGAN dataset using an autoencoder scheme, and b) an architecture of multi-task learning for facial weakness classification.

V. EXPERIMENTAL SETUP AND RESULTS

A. EXPERIMENTAL SETUP

1) DATA LOADER FOR PRE-TRAINING PHASE

A stratified random sampling approach was employed to pre-train the encoder and decoder of the autoencoder model for learning neutral and smiling facial expressions. A total of 14,400 pairs of FaceGAN images were divided into 80% training set and 20% validation set (Fig. 5). The validation set was used to determine a set of hyperparameters that would optimize the performance of the encoder and decoder of the autoencoder model.

2) DATA LOADER FOR MODEL EVALUATION

After training and validating the encoder and decoder of the autoencoder model, we use both encoder and decoder for multi-task learning to create a classification model. To reduce bias and enhance model performance, a stratified k-fold cross-validation (with $k = 4$) is employed due to the limited amount of real subjects' data. The FaceGAN dataset and $k - 1$ fold from the real subjects' dataset are utilized for training and validation of the proposed model in proportions of 80% and 20%, respectively. It should be noted that StyleCLIP and OverLORD GAN augmentation

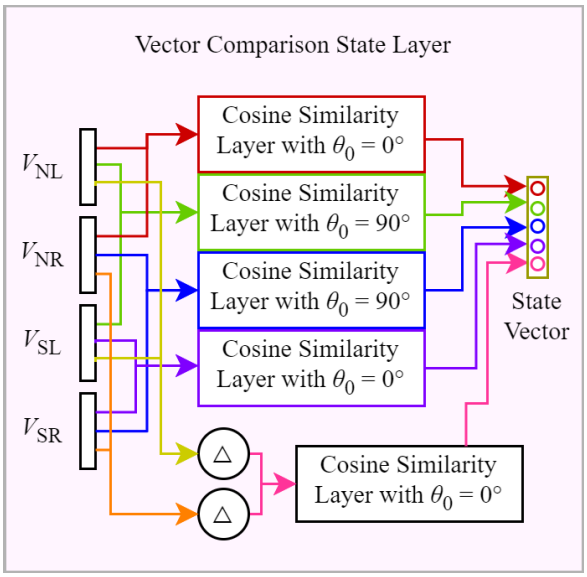

FIGURE 4. Our proposed vector comparison state layer to detect the degree of difference between each part of a facial image.

are applied only during the training stage, while the validating and testing stages were done using original images

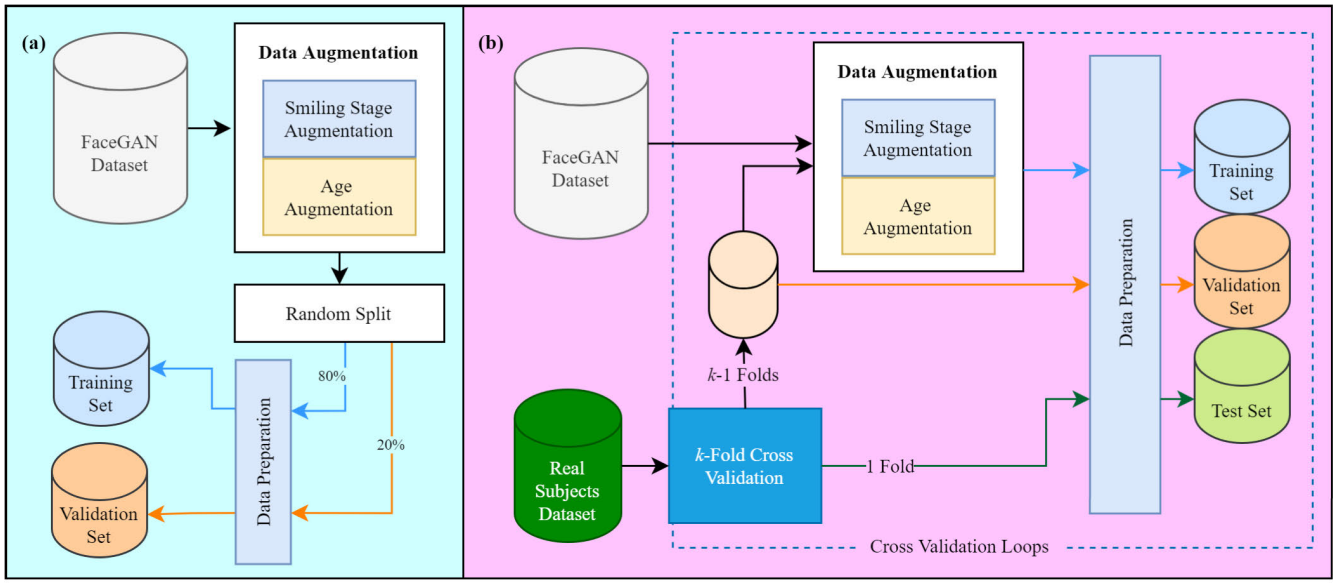

FIGURE 5. The data division used for training the model consists of (a) a data loader for pre-training phase and (b) a data loader for model evaluation.

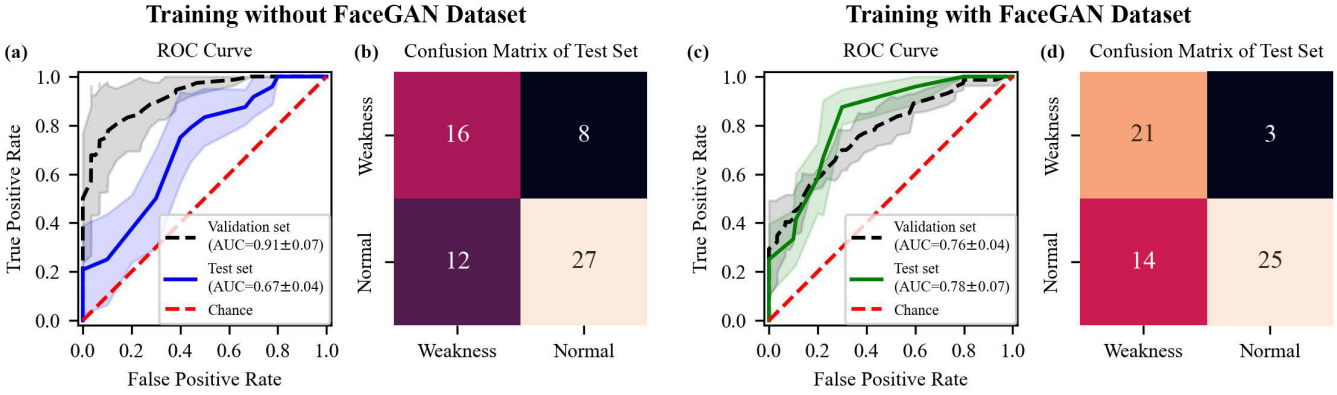

FIGURE 6. Model performance comparison: without FaceGAN dataset vs. with FaceGAN dataset. (a) & (c) present the mean and standard deviation of AUC values for validation and independent testing datasets across four-fold cross-validation, while (b) & (d) display the corresponding confusion matrices from independent testing datasets.

from the real subjects' dataset. The last remaining fold of the real subjects' dataset, without any augmentation techniques, is used to evaluate the performance of the model, as shown in Fig. 5 b). This procedure is repeated four times until all data from the real subjects' dataset have been evaluated.

B. EXPERIMENTAL RESULTS

This study aims to investigate the efficacy of using a FaceGAN dataset to increase the generalizability of a model for distinguishing between normal stages (normal control participants and stroke patients without facial weakness) and facial weakness stage (stroke patients with slight and profound facial weakness). The results are divided into two sections, one examining the classification results of a model trained on data without the FaceGAN dataset, and the other examining the classification results of a model trained on data with the FaceGAN dataset.

C. CLASSIFICATION PREDICTION OF MODEL WITHOUT FaceGAN DATASET

This subsection presents a model trained with data from real subjects dataset and evaluates its performance through four-fold cross validation. Figure 6 (a) depicts the mean AUC and standard deviation of AUC for this model as 0.91 and 0.07, respectively, indicating high performance on training and validating data. However, when tested with an independent dataset, the model's performance dropped to a mean and standard deviation of AUC of 0.67 and 0.04, respectively. The confusion matrix of four rounds revealed that the model has a sensitivity, specificity, F1-score, and Cohen's kappa of 66.67%, 69.23%, 61.54%, and 34.78%, respectively, as shown in Fig. 6 (b). The results demonstrate the model's potential for achieving high performance on training and validating data. However, its performance on independent testing datasets is relatively low, possibly due to the limited available data causing overfitting resulting in minimal

agreement between the predicted and actual labels [42]. Further research is needed to improve the reliability of the model's results on independent datasets.

D. CLASSIFICATION PREDICTION OF MODEL WITH FaceGAN DATASET

The results in this subsection consist of both data from the real subjects dataset and its augmented data used for training, as well as the FaceGAN dataset. Performance from four-fold cross validation using the AUC curve revealed that the model with FaceGAN data had a lower average AUC than the model with only the real subjects dataset for training, with an average and standard deviation of AUC of 0.76 and 0.04, respectively, as shown in Fig. 6 (c). This is likely due to the fact that the training and validation from the previous section was overfitted when both the real subjects and FaceGAN data were used. However, the model with FaceGAN dataset was slightly better than not using the FaceGAN dataset as the average AUC of the testing set is closer to that of the validation dataset with a higher Cohen's kappa level of agreement. The sum of the confusion matrix of four rounds found that the model has a sensitivity, specificity, F1-score, and Cohen's kappa of 87.50%, 64.10%, 71.19%, and 47.42%, as shown in Fig. 6 (d). These results suggest that the model is a promising clinical tool with further development and optimization.

VI. DISCUSSION AND CONCLUSION

Facial paralysis or asymmetry, and difficulty forming facial expressions are indicators of stroke. AI-based stroke detection is optimized when the patient is asked to display both neutral and smiling facial expressions. This process enables a more thorough assessment of facial muscles, as the presence of paralysis and asymmetry becomes more obvious. Furthermore, with faster stroke recognition, prompt treatment can be administered without haste to reduce risks of serious complications or death.

In our work, we utilized a pair of neutral and smiling facial images to classify between normal (including both normal control participants and stroke patients without facial weakness) and facial weakness (including stroke patients with slight and profound facial weakness). Our work takes advantage of the dynamic facial features using only two frames, while other studies incorporate an entire video. Hence, our work cannot be compared with other video classifications. The model generated a very good performance when trained using only real-subject dataset. However, when applied to an anonymous dataset, the results gradually dropped and did not align with the training results. When our proposed techniques were applied together for data augmentation, the performance of validating and testing dataset with anonymous data aligned with each other and produced better outcomes.

This work not only relies heavily on data augmentation techniques, but also utilizes transformation matrices, facial landmarks detection algorithms, and Delaunay triangulation to correct geometric distortions of images. The results support the proposed vector comparison state layer for comparing similarities between neutral and smiling facial images. Moreover, a multi-task learning of multiple losses functions is designed to achieve multiple objectives, resulting in better generalization, faster convergence, improved accuracy even with limited resources. Although this idea is supported by many existing literature, its conclusion has yet to be examined in this study.

The main drawback of this work is the limited availability of real subject datasets. To ensure a balanced dataset and provide enough data to accurately evaluate the model's performance, we decided to include both normal control participants and stroke participants with no facial weakness in the normal group without any preprocessing. Unfortunately, this decision is responsible for the decreased performance seen in both the validation and testing sessions. To ensure thorough evaluation of the model's performance, it was necessary to maintain a sufficient amount of datasets.

Apart from assessment of facial weakness, AI-based techniques can also be used to detect limb weakness and speech impairment, thus providing a more comprehensive approach for detecting stroke. In a previous study [3], arm weakness detection helped in diagnosis of stroke and achieved sensitivity and specificity of 94.8% and 84.1%, respectively. While the sensitivity was high, the specificity was relatively low. This indicates that one modality alone may not adequately detect stroke. Hence, combining several modalities can enhance accuracy of stroke detection leading to better outcomes.

We acknowledge that using progressive growing FGANs for data augmentation could potentially improve our model's performance. However, due to limited resources and number of real subjects, we decided to use the synthesized FaceGAN dataset to create a diverse distribution of age groups, skin color, ethnicity, and appearance. In the future, we plan to collect more data to increase the accuracy of facial weakness screening. Furthermore, we plan to incorporate progressive FGANs like in PanoHead [12], to enhance our model so that it can be used on different face angles and 3D face models. Model evaluation can be improved by adding more performance metrics such as FrB)chet Inception Distance (FID) and Inception Score (IS) [43].

Additionally, we will create an application that combines the model of our previous research on arm weakness detection [3] and this facial weakness detection program to create a more comprehensive stroke screening tool. This application aims for faster stroke detection and enhanced patient triage to improve stroke treatment outcomes. Ultimately, we hope to create an accessible tool that can be used by the general population monitoring and early detection of stroke.

ACKNOWLEDGMENT

The authors would like to thank all subjects who participated in the data collection.

REFERENCES

[1] V. L. Feigin, M. Brainin, B. Norrving, S. Martins, R. L. Sacco, W. Hacke, M. Fisher, J. Pandian, and P. Lindsay, "World stroke organization (WSO): Global stroke fact sheet 2022," *Int. J. Stroke*, vol. 17, no. 1, pp. 18–29, Jan. 2022.

[2] Y. Zhuang, M. McDonald, O. Uribe, X. Yin, D. Parikh, A. M. Southerland, and G. K. Rohde, "Facial weakness analysis and quantification of static images," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2260–2267, Aug. 2020.

[3] P. Phienphanich, N. Tankongchamruskul, W. Akarathanawat, A. Chutinet, R. Nimnual, C. Tantibundhit, and N. C. Suwanwela, "Stroke screening feature selection for arm weakness using a mobile application," *IEEE Access*, vol. 8, pp. 170898–170914, 2020.

[4] Y. Zhuang, M. M. McDonald, C. M. Aldridge, M. A. Hassan, O. Uribe, D. Arteaga, A. M. Southerland, and G. K. Rohde, "Video-based facial weakness analysis," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 9, pp. 2698–2705, Sep. 2021.

[5] K.-Y. Lee, C.-C. Liu, D. Y.-T. Chen, C.-L. Weng, H.-W. Chiu, and C.-H. Chiang, "Automatic detection and vascular territory classification of hyperacute staged ischemic stroke on diffusion weighted image using convolutional neural networks," *Sci. Rep.*, vol. 13, no. 1, p. 404, Jan. 2023.

[6] P. Phienphanich, N. Tankongchamruskul, W. Akarathanawat, A. Chutinet, R. Nimnual, C. Tantibundhit, and N. C. Suwanwela, "Automatic stroke screening on mobile application: Features of gyroscope and accelerometer for arm factor in FAST," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 4225–4228.

[7] H. P. Adams Jr., G. D. Zoppo, M. J. Alberts, D. L. Bhatt, L. Brass, A. Furlan, R. L. Grubb, R. T. Higashida, E. C. Jauch, C. Kidwell, P. D. Lyden, L. B. Morgenstern, A. I. Qureshi, R. H. Rosenwasser, P. A. Scott, and E. F. M. Wijdicks, "Guidelines for the early management of adults with ischemic stroke," *Stroke*, vol. 38, no. 5, pp. 1655–1711, 2007. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.107.181486

[8] E. S. Brandler, M. Sharma, R. H. Sinert, and S. R. Levine, "Prehospital stroke scales in urban environments: A systematic review," *Neurology*, vol. 82, no. 24, pp. 2241–2249, Jun. 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24850487/

[9] I. Induruwa, N. Holland, R. Gregory, and K. Khadjooi, "The impact of misdiagnosing Bell's palsy as acute stroke," *Clin. Med.*, vol. 19, no. 6, pp. 494–498, Nov. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6899254/

[10] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2085–2094.

[11] A. Gabbay and Y. Hoshen, "Scaling-up disentanglement for image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6783–6792.

[12] S. An, H. Xu, Y. Shi, G. Song, U. Ogras, and L. Luo, "PanoHead: Geometry-aware 3D full-head synthesis in 360°," 2023, *arXiv:2303.13071*.

[13] G. S. Parra-Dominguez, C. H. Garcia-Capulin, and R. E. Sanchez-Yanez, "Automatic facial palsy diagnosis as a classification problem using regional information extracted from a photograph," *Diagnostics*, vol. 12, no. 7, p. 1528, Jun. 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/35885434

[14] Z. Guo, M. Shen, L. Duan, Y. Zhou, J. Xiang, H. Ding, S. Chen, O. Deussen, and G. Dan, "Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 135–138.

[15] M. Rajnoha, J. Mekyska, R. Burget, I. Eliasova, M. Kostalova, and I. Rektorova, "Towards identification of hypomimia in Parkinson's disease based on face recognition methods," in *Proc. 10th Int. Congr. Ultra Modern Telecommun. Control Syst. Workshops (ICUMT)*, Nov. 2018, pp. 1–4.

[16] T. H. Ngo, M. Seo, N. Matsushiro, and Y. Chen, "Quantitative analysis of facial paralysis based on filters of concentric modulation," in *Proc. 12th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, Aug. 2015, pp. 1758–1763.

[17] X. Hou, Y. Zhang, Y. Wang, X. Wang, J. Zhao, X. Zhu, and J. Su, "A markerless 2D video, facial feature recognition–based, artificial intelligence model to assist with screening for Parkinson disease: Development and usability study," *J. Med. Internet Res.*, vol. 23, no. 11, Nov. 2021, Art. no. e29554. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/34806994

[18] B. Jin, Y. Qu, L. Zhang, and Z. Gao, "Diagnosing Parkinson disease through facial expression recognition: Video analysis," *J. Med. Internet Res.*, vol. 22, no. 7, Jul. 2020, Art. no. e18697. [Online]. Available: https://www.jmir.org/2020/7/e18697

[19] A. Bandini, S. Orlandi, H. J. Escalante, F. Giovannelli, M. Cincotta, C. A. Reyes-Garcia, P. Vanni, G. Zaccara, and C. Manfredi, "Analysis of facial expressions in Parkinson's disease through video-based automatic methods," *J. Neurosci. Methods*, vol. 281, pp. 7–20, Apr. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027017300481

[20] D. L. Guarin, Y. Yunusova, B. Taati, J. R. Dusseldorp, S. Mohan, J. Tavares, M. M. van Veen, E. Fortier, T. A. Hadlock, and N. Jowett, "Toward an automatic system for computer-aided assessment in facial palsy," *Facial Plastic Surg. Aesthetic Med.*, vol. 22, no. 1, pp. 42–49, Feb. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7362997/

[21] G. Storey, R. Jiang, S. Keogh, A. Bouridane, and C. Li, "3DPalsyNet: A facial palsy grading and motion recognition framework using fully 3D convolutional neural networks," *IEEE Access*, vol. 7, pp. 121655–121664, 2019.

[22] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[23] Y. Gizatdinova and V. Surakka, "Feature-based detection of facial landmarks from neutral and expressive facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 135–139, Jan. 2006.

[24] Y. Guo, G. Zhao, and M. Pietikäinen, "Dynamic facial expression recognition with atlas construction and sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1977–1992, May 2016.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[26] C. M. Aldridge, M. M. McDonald, M. Wruble, Y. Zhuang, O. Uribe, T. L. Mcmurry, I. Lin, H. Pitchford, B. J. Schneider, W. A. Dalrymple, J. F. Carrera, S. Chapman, B. B. Worrall, G. K. Rohde, and A. M. Southerland, "Human vs. machine learning based detection of facial weakness using video analysis," *Frontiers Neurol.*, vol. 13, p. 878, Jul. 2022.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[28] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, "Generative adversarial networks for face generation: A survey," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–37, May 2023.

[29] Generated Media. (2022). *Image Datasets for Machine Learning | Generated.Photos*. Accessed: Apr. 25, 2022. [Online]. Available: https://generated.photos/datasets

[30] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2014, pp. 94–108.

[31] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 115–142, Feb. 2019.

[32] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5202–5211.

[33] W. K. Pratt, *Digital Image Processing: PIKS Scientific Inside*, vol. 4. Hoboken, NJ, USA: Wiley, 2007.

[34] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, Mar. 2017.

[35] Y. Zhong, J. Chen, and B. Huang, "Toward end-to-end face recognition through alignment learning," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1213–1217, Aug. 2017.

[36] D. T. Lee and B. J. Schachter, "Two algorithms for constructing a Delaunay triangulation," *Int. J. Comput. Inf. Sci.*, vol. 9, no. 3, pp. 219–242, Jun. 1980.

[37] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.

[38] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 646–661.
[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[40] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.
[41] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.
[42] M. L. McHugh, "Interrater reliability: The Kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.
[43] M. J. Chong and D. Forsyth, "Effectively unbiased FID and inception score and where to find them," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6070–6079.

PHONGPHAN PHIENPHANICH (Student Member, IEEE) received the B.E. degree (Hons.) in computer engineering from the Suranaree University of Technology, Nakhon Ratchasima, Thailand, in 2009, and the M.E. degree in electrical engineering from Thammasat University, Bangkok, Thailand, in 2012, where he is currently pursuing the Ph.D. degree in computer engineering. From 2010 to 2012, he was a Co-Researcher with the National Electronics and Computer Technology Center (NECTEC), Thailand. His research interests include signal and speech processing, pattern recognition, and machine learning.

NICHAPA LERTHIRUNVIBUL received the medical degree (Hons.) from Thammasat University, Bangkok, Thailand. She is currently a Researcher with the Center of Excellence in Intelligent Informatics, Speech and Language Technology, and Service Innovation (CILS), Thammasat University. Her research interest includes the use of artificial intelligence in healthcare and medicine.

EKABHAT CHARNNARONG is currently pursuing the Patumwan Demonstration School, Bangkok, Thailand. His research interests include signal and speech processing, pattern recognition, and machine learning.

ADIREK MUNTHULI received the B.E. degree in computer engineering and the M.E. degree in electrical and computer engineering from Thammasat University, Bangkok, Thailand, where he is currently pursuing the Ph.D. degree in computer engineering. His research interests include natural language processing, speech and signal processing, and machine learning.

CHARTURONG TANTIBUNDHIT (Member, IEEE) received the B.E. degree in electrical engineering from Kasetsart University, Bangkok, Thailand, in 1996, and the M.S. degree in information science and the Ph.D. degree in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, USA, in 2001 and 2006, respectively. Since 2006, he has been with Thammasat University, Thailand, where he is currently an Associate Professor with the Department of Electrical and Computer Engineering and the Head of the Speech and Language Technology Cluster, Center of Excellence in Intelligence Informatics, Speech and Language Technology, and Service Innovation (CILS). From 2007 to 2008, he was a Postdoctoral Researcher with the Signal Processing and Speech Communication Laboratory (SPSC), Graz University of Technology, Graz, Austria. His research interests include handcrafted machine learning and deep learning in medicine, biomedical signal processing, and speech processing. He was the IEEE ICASSP Student Paper Contest Winner, in 2006. He led a team that won the Grand Prix from the 45th International Exhibition of Inventions of Geneva, in 2017.

NIJASRI C. SUWANWELA received the M.D. degree (Hons.) from Chulalongkorn University, Bangkok, Thailand, in 1989. She had her residency training with King Chulalongkorn Memorial Hospital, Bangkok, in 1993. She was later awarded by the King Anandamahidol Foundation to study as a fellow in cerebrovascular diseases with the Massachusetts General Hospital, Boston, MA, USA, in 1996. After returning to Thailand, she pioneered the thrombolysis use and neurosonology in the country. She is currently the President of the Neurological Society of Thailand and a Professor and the Director of the Chulalongkorn Comprehensive Stroke Center, Bangkok. She is also the former Head of the Neurology Division, Chulalongkorn University, the Vice President of the Neurological Society of Thailand and the Thai Stroke Society, and the President of the Asian Stroke Advisory Panel. She has published more than 50 peer-reviewed articles and many book chapters.