Improving Distinguishability of Photoplethysmography in Emotion Recognition Using Deep Convolutional Generative Adversarial Networks

SUNG-NIEN YU, SHAO-WEI WANG, AND YU PING CHANG

Department of Electrical Engineering, National Chung Cheng University, Minxiong, Chiayi County 621301, Taiwan
Center for Innovative Research on Aging Society (CIRAS), National Chung Cheng University, Minxiong, Chiayi County 621301, Taiwan

Corresponding author: Sung-Nien Yu (iecstny@ccu.edu.tw)

This work was supported in part by the Center for Innovative Research on Aging Society (CIRAS) from the Featured Areas Research Center Program within the Framework of the Higher Education Sprout Project by the Ministry of Education (MOE), Taiwan; and in part by the Ministry of Science and Technology (MOST), Taiwan, through the Pervasive Artificial Intelligence Research (PAIR) Laboratories under Contract MOST 108-2221-E-194-034-MY3, Contract MOST 108-2634-F-037-001, and Contract 109-2634-F-037-001.

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 Kaohsiung Medical University Hospital under Application No. KMUHIRB-E(I)-20180327.

ABSTRACT We propose an emotion recognition framework based on ResNet, bidirectional long- and short-term memory (BiLSTM) modules, and data augmentation using a ResNet deep convolutional generative adversarial network (DCGAN) with photoplethysmography (PPG) signals as input. The emotions identified in this study were classified into two classes (positive and negative) and four classes (neutral, angry, happy, and sad). The framework achieved high recognition rates of 90.34% and 86.32% in two- and four-class emotion recognition tasks, respectively, outperforming other representative methods. Moreover, we show that the ResNet DCGAN module can synthesize samples that do not just look like those in the training set, but also capture discriminative features of the different classes. The distinguishability of the classes was enhanced when these synthetic samples were added to the original samples, which in turn improved the test accuracy of the model when trained using these mixed samples. This effect was evaluated using various quantitative and qualitative methods, including the inception score (IS), Fréchet inception distance (FID), GAN quality index (GQI), linear discriminant analysis (LDA), and Mahalanobis distance (MD).

INDEX TERMS Augmentation, DCGAN, emotion recognition, PPG.

I. INTRODUCTION

The study of coronary artery heart disease (CAD) has been a focus of clinical health and engineering research. In Taiwan, heart disease ranks second among the top ten causes of death. And among heart diseases, 52% deaths are due to CAD, according to a report by the Ministry of Health and Welfare. Besides physiological risk factors, psychosocial factors such as strong emotions, including anger, hostility, anxiety, and depression, have also been identified as leading causes of CAD. Rosengren et al. [1] confirmed that psychosocial risk factors cause long-term and repeated psychophysiological responses that affect the autonomic nervous system (ANS) and neuroendocrine system, leading to CAD. With the development of AI techniques, it is hoped that negative emotions can be identified in the early stages so that appropriate psychological intervention can be activated.

In automated emotion recognition, many studies have focused on visual or auditory signals such as facial expressions [2], [3], [4], [5], [6], [7], [8], [9], [10].
text [10], [11], [12], and gestures [13], [14], [15]. This is intuitive because these signals are most often involved in human-human interactions. More recently, AI systems based on physiological signals, such as photoplethysmography (PPG) [16], electrocardiography (ECG) [17], galvanic skin response (GSR) [18], electroencephalography (EEG) [19], [20], electromyography (EMG) [21], and respiratory pattern (RSP) [22] have been studied, and these physiological changes that correlate with different emotional states are often more truthful than externally expressed [23].

PPG in particular can be conveniently measured either directly using a small device attached to the finger or remotely through facial video [24]. PPG is a low-cost and noninvasive optical technique that measures blood volume changes at the skin surface. The pulsatile component of PPG correlates strongly with each heartbeat and so has often been used as an alternative to ECG to provide information about the cardiac system [25]. More importantly, the frequency-domain features of short-time PPG have been shown to respond significantly to emotion changes [26]. Therefore, we selected PPG as the input signal and converted it into time-frequency domain spectrogram images to classify the short-term emotions.

Various techniques have been developed for automated emotion recognition, and the following research motivated this study. Lee et al. [27] used single-pulse PPG signals from the DEAP database to identify two types of emotion. They employed a one-dimensional convolutional neural network (1D CNN) to extract PPG features for classification and obtained satisfactory results. Etienne et al. [28] used the EMOCAP speech dataset and designed an architecture with multiple convolutional layers to extract high-level features, and a bidirectional long- and short-term memory network (BiLSTM) to model long-term dependencies. To further address the overfitting problem due to small datasets, as is often the case in emotion classification problems, Eskimez et al. [29] designed data augmentation based on a deep convolutional generative adversarial network (DCGAN) to generate synthetic speech spectrogram images to classify emotions and showed that using purely GAN-generated data to train the model could also obtain comparable results.

We combined the advantages of these studies and proposed an emotion recognition framework based on PPG signals using ResNet BiLSTM architecture and ResNet DCGAN data augmentation. The ResNet module extracts complex and detailed features from the spectrogram, whereas the BiLSTM module captures the long-term dependencies between the features extracted at different time points. The ResNet BiLSTM architecture was shown to achieve good emotion classification results in our previous study [30].

In this study, we extended our earlier work and addressed the generalization problem by employing the DCGAN technique. Specifically, this paper proposes ResNet DCGAN data augmentation, which generates synthetic spectrograms that not only simulate the distribution of the original training samples, but also emphasize the discriminative features of the different classes when synthesizing new samples, thereby improving the final classification accuracy. Five different metrics were used to evaluate the DCGAN module: the inception score (IS), Fréchet inception distance (FID), GAN quality index (GQI), linear discriminant analysis (LDA), and Mahalanobis distance (MD).

The remainder of this paper is organized as follows. Section 2 briefly summarizes ResNet, BiLSTM, and the results of our previous study using this architecture for emotion recognition. Section 3 describes the data augmentation method proposed in this study using DCGAN and the five evaluative methods mentioned above. Section 4 outlines the proposed emotion recognition framework, as well as the data collection, preprocessing, and experimental procedures. The results and discussions on sample generation and emotion classification, together with the ablation study and a comparison with relevant studies, are detailed in Section 5. We conclude the paper in Section 6.

The contributions of this study are twofold:

1. We demonstrated the efficacy of the proposed emotion recognition framework which employed the ResNet BiLSTM architecture and ResNet DCGAN data augmentation using PPG spectrograms as the input. High emotion recognition accuracy was achieved.
2. Using various quantitative and qualitative metrics, we demonstrate the ability of the ResNet DCGAN module to generate samples that go beyond the training sample distribution, but also identify discriminative features of each class, and thus enhance the distinguishability of classes when mixed with the original samples.

II. RELEVANT METHODOLOGICAL BACKGROUND

A. RESNET MODULE

One of the most notable contributions of the ResNet architecture proposed by He et al. [31] in 2015 is the shortcut connection, which effectively overcomes the vanishing gradient and degradation problems of deep neural networks with many layers. These shortcut connections add the input feature map to the output after the next few layers, which forces the in-between layers to learn the residual function, such that the layers either improve or retain the current feature map. Thus, regardless of the depth of the model, only the layers that contribute to improving the model are learned, and the remaining layers can be effectively reduced to identity mapping, thereby mitigating the degradation problem owing to the large number of layers in the network. This architecture has been successfully implemented in various image-related problems; therefore, it was selected as the main feature-extraction module in this study.

B. BILSTM MODULE

BiLSTM is a recently developed architecture under the category of recurrent neural networks (RNN) that aims to model dependencies among data points of a sequence of data,
often temporal in nature. However, traditional RNN fail to capture long-term dependencies due to vanishing gradient problem, which led researchers Hochreiter and Schmidhuber [32] to develop the long- and short-term memory (LSTM) architecture in 1997. The LSTM incorporates gated memory, including the “input gate,” “output gate,” and “forget gate” which effectively retains in its internal state memory useful information from previously encountered tokens to be considered when future tokens arrive. However, this only modeled dependencies with tokens encountered earlier, not later than the current token, which is a limitation in many situations. Schuster et al. [33] in 1997 proposed a bidirectional recurrent neural network that ran the input sequence in both forward and backward directions. The same was applied to the LSTM architecture to form BiLSTM, which was shown to perform better than its unidirectional version [34].

We combined the above-mentioned ResNet and BiLSTM modules to capture the complex features within the PPG spectrogram images as well as their temporal dynamics. The resulting ResNet BiLSTM model is shown in Table 1. Experimental results from our preliminary study [30] showed that the proposed ResNet BiLSTM model was capable of recognizing two- and four-class emotions with accuracies of 89.15% and 84.70%, respectively.

III. DATA AUGMENTATION USING DCGAN AND ITS EVALUATION
A. DCGAN
The generative adversarial network (GAN) architecture was first invented by Goodfellow et al. [35] in 2014 and originally used multilayer perceptron in the generator and discriminator modules. Radford et al. [36] adapted the framework in 2015 and proposed deep convolutional generative adversarial networks (DCGAN), which employed convolution and fractionally strided convolution layers (some call deconvolution) in place of multilayer perceptron in the discriminator and generator modules. The generator module samples values from the Gaussian distribution as input to its latent variables and generates faithful synthetic samples that resemble the originals, whereas the discriminator module learns to distinguish fake from real ones. Through this two-player game, the two modules iteratively improve each other, eventually producing a generator module that is effective at generating samples that capture the essence of the original samples.

This study compares two different DCGAN architectures. The first model is the standard DCGAN proposed by Radford et al. [36], and is outlined in Table 2. The second architecture incorporates ResNet into the DCGAN using shortcut connections to mitigate vanishing gradient and degradation problems, as detailed in Table 3.

Latent variables of length 100 were sampled from a Gaussian distribution as inputs to the generator module to generate synthetic samples. These generated samples were then mixed with true samples to be fed into the discriminator module of the DCGAN to differentiate true from fake. In each iteration, the parameters of the modules were updated, as shown in Fig. 1.

The objective function of the DCGAN model is defined in (1) below.

\[ V(D, G) = E_x [\log(D(x))] + E_z [\log(1 - D(G(z)))]. \]  

where \( D(x) \) is the discriminator’s probability of predicting real sample \( x \) as the true sample, and \( D(G(z)) \) refers to the probability of predicting the synthetic sample \( G(z) \) generated by the generator module as the true sample. \( z \) is the input to the generator module, sampled from a Gaussian distribution. The

| TABLE 1. ResNet BiLSTM architecture and parameter settings. |
| Layer name | Parameters |
|------------|------------|
| Input layer | [64 x 64 x 1] |
| Conv layer | [10 x 10, 64, stride2] |
| Pooling layer | Max Pooling |
| Conv layer | [3 x 3, 64, stride2] x 2 |
| Conv layer | [3 x 3, 128, stride2] x 2 |
| Conv layer | [3 x 3, 256, stride2] x 1 |
| Pooling layer | Global Average Pooling |
| Flatten layer | |
| BiLSTM layer | Hidden unit: 128 |
| BiLSTM layer | Output mode: sequence |
| Activation Layer | Tanh |

| TABLE 2. Original DCGAN parameter settings. |
| Layer name | Parameter |
|------------|------------|
| Input layer | [64 x 64 x 1] |
| Conv layer | [4 x 4, 16, stride2] |
| Conv layer | [4 x 4, 32, stride2] |
| Conv layer | [4 x 4, 64, stride2] |
| Conv layer | [4 x 4, 128, stride2] |
| Conv layer | [4 x 4, 256, stride2] |
| Conv layer | [4 x 4, 1, stride1] |
| Activation Layer | Tanh |

| TABLE 3. ResNet DCGAN parameter settings. |
| Layer name | Parameter |
|------------|------------|
| Input layer | [64 x 64 x 1] |
| Conv layer | [4 x 4, 16, stride2] |
| Conv layer | [4 x 4, 32, stride2] |
| Conv layer | [4 x 4, 64, stride1] |
| Conv layer | [4 x 4, 128, stride1] |
| Conv layer | [4 x 4, 256, stride1] |
| Conv layer | [4 x 4, 1, stride4] |
| Activation Layer | Tanh |
generator module of the DCGAN was trained to minimize the objective function, whereas the discriminator maximized it.

In this study, four types of emotion were recognized: neutral, angry, happy, and sad. We intend to augment each set of samples to double the original number, and employ five metrics to evaluate the performance of the generator modules. The metrics include the inception score (IS), Fréchet inception distance (FID), GAN quality index (GQI), linear discriminant analysis (LDA), and Mahalanobis distance (MD).

B. EVALUATING THE GENERATOR MODULE (IS, FID, GQI, LDA, MD)

Evaluation methods to test the effectiveness of DCGAN in generating convincing samples have been widely explored. Although a straightforward method involves experts looking at the generated samples to evaluate their credibility, it can be polluted by subjective opinions. Therefore, in this study we considered five different objective metrics: inception score (IS) [37], Fréchet inception distance (FID) [38], GAN quality index (GQI) [39], linear discriminant analysis (LDA) [40], and Mahalanobis distance (MD) [41].

1) INCEPTION SCORE (IS)

The inception score was first proposed by Salimans et al. [37]. The idea is to use a pretrained InceptionNet-V3 as the classifier and grade the quality of the synthetic samples produced by the generator module of the DCGAN using the prediction probability distribution. Barratt et al. [42] advised in their note on the inception score that InceptionNet-V3 was pretrained using images from ImageNet; therefore, the inception score based on such a pretrained InceptionNet-V3 can only be used for samples similar to those in ImageNet. In our case, the samples were medical spectrograms, which are quite different from the natural images in ImageNet. To mitigate this problem, we substituted the InceptionNet-V3 model with the ResNet model trained on our dataset to determine the quality of the generator. The score is calculated in two parts as follows:

\[ IS = \exp \left( \frac{1}{N} \sum_{i=1}^{N} D_{KL} \left( p \left( y \mid x^{(i)} \right) \| p \left( y \right) \right) \right) \]  

where the values in each dimension represent the conditional probability \( p \left( y \mid x^{(i)} \right) \) of a given sample belonging to a certain class. Ideally, a good-generated sample should have a high probability of belonging to one class and a low probability of belonging to all other classes.

2) FRÉCHET INCEPTION DISTANCE (FID)

The Fréchet inception distance was proposed by Heusel et al. [38]. Compared with IS, FID mitigates the problem of InceptionNet-V3 being trained on ImageNet which is different from the actual task at hand. The FID uses the pretrained InceptionNet-V3 to extract features from synthetic samples and compares their distribution with the features extracted from the true samples. Again, we replaced InceptionNet-V3 with our own ResNet model, which was trained using the samples from this study. The FID was calculated as follows:

\[ \text{FID} = \left\| \mu_r - \mu_g \right\|^2 + \text{Tr} \left( \Sigma_r + \Sigma_g - 2 \left( \Sigma_r \Sigma_g \right)^{1/2} \right) \]  

In this study, the 256-dimension vector output after the average pooling operation of the ResNet model was considered as the feature vector. \( \mu_r \) and \( \mu_g \) represent the mean feature values, whereas \( \Sigma_r \) and \( \Sigma_g \) represent the covariance matrices of the real and generated samples, respectively. The FID ranges from 0 to infinity; the smaller the value, the better is the generator module.

FIGURE 1. Block diagram of DCGAN.
3) GAN QUALITY INDEX (GQI)
Ye et al. proposed the GAN quality index [39]. The idea is that if the generated samples have a similar distribution to the original samples, then training a classifier from the same number of either the generated or original samples should yield a similar prediction accuracy. The calculation is defined in (6).

\[
GQI = \frac{Acc(C_{GAN})}{Acc(C_{real})} \times 100
\]  

(6)

Consider a dataset that contains \( N \) distinct categories, a classifier \( C_{real} \) can be trained using this dataset of real samples to obtain an accuracy \( Acc(C_{real}) \) when evaluated against a test dataset. Subsequently, a generative model was trained using the same dataset. Synthetic samples of the same amount were then generated and fed into the classifier trained earlier, \( C_{real} \), to obtain the categories to which these generated samples should belong. A probability threshold can be set to filter out the samples that do not belong to any class. A new classifier, \( C_{GAN} \), was trained using the newly obtained dataset of generated samples. The accuracy achieved by evaluation against the test dataset \( Acc(C_{GAN}) \) was obtained. Using these two accuracy values, the GQI value can be calculated using (6). The GQI values range from 1 to 100; the larger the value, the closer the distribution of the generated samples to the original real samples.

4) LINEAR DISCRIMINANT ANALYSIS (LDA)
In contrast to the three evaluation methods outlined earlier, linear discriminant analysis (LDA) allows the evaluation of generated samples using a visual method instead of a score. LDA is a supervised learning method, the main usage of which is dimensionality reduction and classification [40]. LDA projects the samples onto a lower dimension and calculates the within-class and between-class scatter of the projected data. To maximize the separation between different classes, LDA maximizes the Fisher criterion to minimize the within-class differences and maximize the between-class differences. This study employed LDA to reduce the dimensionality of the 256-dimensional features extracted from the ResNet model to 2-dimensions so that class distinguishability can be visually observed from a 2D plot. The detailed LDA procedure is as follows.

We assume a binary classification problem and consider a dataset with 200 samples, with 100 samples in each set, namely, sets 1 and 2:

1. Calculate the mean of the two sets \( \mu_1 \) and \( \mu_2 \), and then the mean of all samples \( \mu_3 \) as given in (7), where \( p_1 \) and \( p_2 \) are a priori probabilities.

\[
\begin{align*}
\mu_3 &= p_1 \times \mu_1 + p_2 \times \mu_2 \\
&= \text{mean of all samples}
\end{align*}
\]

(7)

2. The within-class scatter \( S_w \) and between-class scatter \( S_B \) were then calculated using (8) and (9), respectively, where \( j \) is the index for the classes. The covariance for the samples of each class can then be calculated using (10).

\[
S_w = \sum_j p_j \times (\text{cov}_j)
\]

(8)

\[
S_B = \sum_j (\mu_j - \mu_3) \times (\mu_j - \mu_3)^T
\]

(9)

\[
\text{cov}_j = (x_j - \mu_j)(x_j - \mu_j)^T
\]

(10)

The calculation is defined in (11).

\[
\text{criterion}_j = \text{inv}(S_w) \times S_B
\]

(11)

5) MAHALANOBIS DISTANCE (MD)
The Mahalanobis distance (MD) was proposed by Indian statistician P.C. Mahalanobis [41] and can be regarded as an improved version of Euclidean distance. The Mahalanobis distance corrects the problem of inconsistent scaling in various dimensions when viewed using the Euclidean distance. This method is effective in calculating the similarity between two unknown sample sets. This study employed Mahalanobis distance to evaluate the distance between various categories of samples in the dataset. The calculation process is as follows:

1. Find the mean for each class of samples \( \mu_{Loc} \) and for all samples \( \mu_{Glo} \).
2. Find the covariance matrix for all samples \( \text{cov}_{Glo} \).
3. Find the Mahalanobis distance between the mean of each class and the mean of all samples, as shown in (12), where \( c \) is the index for each class.

\[
MD_c(\mu_{Loc}, \mu_{Glo}) = \sqrt{\text{cov}_{Glo} \cdot (\mu_{Loc} - \mu_{Glo}) \cdot (\mu_{Loc} - \mu_{Glo})^T}
\]

(12)

4. Find the average Mahalanobis distance across all classes \( MD_c \), as shown in (13), where \( nc \) is the number of classes.

\[
MD = \frac{\sum_{c=1}^{nc} MD_c}{nc}
\]

(13)

IV. EXPERIMENTAL PROCEDURE AND DATA COLLECTION
A. EMOTION RECOGNITION FRAMEWORK
A block diagram of the proposed emotion recognition method is shown in Fig. 2. The PPG signals were first preprocessed.
FIGURE 3. The data collecting procedure.

to remove the background noise and baseline fluctuations. The time signals were then converted into spectrograms to train the ResNet DCGAN model and to generate synthetic spectrogram images to augment the original dataset. The mixed spectrogram dataset was then inputted into the proposed ResNet BiLSTM model for training and validation. All computations were coded using Python software.

B. DATA COLLECTION

The dataset used in this study was collected in collaboration with Kaohsiung Medical University Chung-Ho Memorial Hospital (KMUH), Taiwan, under IRB number KMUHIRB-E(I)_20180327. A total of 40 participants (21 men and 19 women) aged 25–90 years, referred by hospital cardiologists, were included in this study.

The experiments were conducted with the assistance of two clinical psychologists from KMUH. The participants were required to complete self-report questionnaires that included demographic characteristics (including age, sex, education, and marital status) before experiments. The data collection procedure is illustrated in Fig. 3.

First, a baseline signal was collected before emotion activation. Then, guided by a clinical psychologist, the participant was asked to state for 180 s and recall for another 180 s a specific emotional experience in alignment with the required emotion category. Subsequently, the participant was allowed to recover and rest for 300 s. This was repeated for four emotional categories (neutral, angry, happy, and sad) using a counterbalanced experimental design. Throughout the entire procedure, a noninvasive PPG sensor (BVP-Flex/Pro) was placed on the participant’s thumb for greatest signal quality [43] and recorded continuously using ProComp InfinitiTM version 6 (Thought Technology Ltd., Montreal, Quebec, Canada). The sensor generates infrared light pulses against the skin surface and measures the amount of reflected light. The PPG sensor was filtered using a 0.1 Hz to 50 Hz bandpass filter and acquired into the device at a sampling rate of 256 samples/s.

An emotional checklist was used to evaluate the specific emotion during the emotional recall tasks, and an emotion rating scale was used to evaluate the emotional intensity (from 1 = not at all to 5 = very) in the past event and during the experimental stages. The data of participants who were unable to recall the target emotions were excluded from the dataset.

In this study, only the 180 s recall period was used as it is more stable than the statement period. In the case of two-class emotion classification, neutral and happy were grouped as the positive category, while angry and sad were grouped as the negative category.

C. PREPROCESSING

As the envisioned use case for this study is in home application, where emotion recognition based on PPG should be performed within a short duration, we set the length of PPG segments to 20 s. The 180 s strips of PPG signals were cropped into 20 s segments with 5 s overlapping to the right. As a result, 1400 segments were obtained for each of the four emotion categories for the 40 participants, totaling 5600 segments. We then converted the 1D PPG segments into spectrograms in the time-frequency domain through continuous wavelet transform (CWT), which has higher precision in time and frequency positioning than the conventional short-time Fourier transform [44]. The continuous wavelet transform is calculated over a finely discretized time-frequency grid, which generates smoothly changing spectrogram images from a 1D signal that can be fed into the image recognition model [45]. The absolute value of the CWT was calculated using the analytic Morse wavelet with a symmetry parameter (gamma) equal to 3 and a time bandwidth product equal to 60. Nearest-neighbor interpolation was then employed to resize the spectrograms to 64 × 64 pixels.

D. CLASS LABELING FOR GENERATED SAMPLES

The synthetic samples generated by the DCGAN model were fed into the ResNet model trained using the original samples. If the output of one of the classes had a probability higher
TABLE 4. Evaluation of quality of generative model.

|                | IS (Mean/Std) | FID  | GQI  |
|----------------|---------------|------|------|
| Original Sample| 3.34 / 0.07   |      |      |
| Standard DCGAN | 3.14 / 0.11   | 5.61 | 49.23|
| ResNet DCGAN   | 3.05 / 0.06   | 14.34| 33.46|

TABLE 5. Evaluating class distinguishability of generative model.

|                | MD  |
|----------------|-----|
| Original Samples| 2.71|
| Standard DCGAN  | 4.57|
| ResNet DCGAN    | 6.52|
| Original Samples + Standard DCGAN | 3.35|
| Original Samples + ResNet DCGAN    | 4.61|

than 0.90, the sample was labeled as the emotion category. This process is illustrated in Fig. 4.

V. RESULTS

A. EVALUATION OF THE GAN MODELS

1) EVALUATING SIMILARITY OF GENERATED SAMPLES TO ORIGINAL SAMPLES

The IS, FID, and GQI were employed to assess the similarity of the generated samples to the original samples. The larger the IS and GQI values and the smaller the FID value, the more similar is the quality and distribution of the resulting sample to the original samples. Table 4 presents the quality assessment of the generative model. As the IS, FID, and GQI values for the standard DCGAN model were superior to those of the ResNet DCGAN model, we considered the synthetic samples generated by the standard DCGAN model to be closer to the original samples than those generated by the ResNet DCGAN model.

2) EVALUATING CLASS DISTINGUISHABILITY OF GENERATED SAMPLES

In addition to assessing the proximity of the generated sample to the original samples, we also assessed the distance between the classes of the generated samples to evaluate their class distinguishability. This is performed by extracting the 256-dimensional features after the generated and original samples pass through the ResNet classifier and calculating their Mahalanobis distance to measure the distance between classes for the generated and original samples. The higher the MD values, the greater the distance between the classes, that is, the higher the distinguishability.

As shown in Table 5, the MD value of the ResNet DCGAN generated samples was significantly higher than those of the original samples and standard DCGAN generated samples. By employing LDA to reduce the dimensionality of the samples to 2-dimensions then plotting the distributions, we obtain Fig. 5. It can be observed that the four classes of the original samples (Fig. 5(a)) were clustered closely together, with samples from different classes overlapping each other such that classification among them was difficult. In contrast, samples generated by the ResNet DCGAN model (Fig. 5(b)) were much easier to distinguish into the four categories. For the samples generated by standard DCGAN (Fig. 5(c)), the distribution appeared similar to that of the original samples.

From the above analysis, we see that the standard DCGAN achieves better results in terms of the IS, FID, and GQI metrics, meaning that its samples are closer in distribution to the original samples, but at the same time, it also inherits the indistinguishability of the original with lower MD values. For the samples generated by the ResNet DCGAN, although less similar to the original samples, its MD is much higher, and from the visualization after dimensionality reduction, it can be seen that the four classes are much more distinguishable. Therefore, the ResNet DCGAN model generated samples with better distinguishability than the standard DCGAN model.

In addition, we calculated the Mahalanobis distance between classes after mixing the generated samples with the original samples to evaluate whether the distinguishability between classes would improve. It can be observed from Table 5 that by mixing the original samples with the samples
generated by ResNet DCGAN, the MD value increased from the original 2.71 to 4.61. Comparatively, if the original samples were mixed with samples generated by the standard DCGAN, the MD value increased only moderately to 3.35. A similar conclusion can be drawn from the distribution plot of the mixed samples, as shown in Fig. 6.

Fig. 6(a) shows the distribution of the four-class original samples; the points are clustered together. In contrast, samples from different emotion categories were more separated when samples generated by ResNet DCGAN were mixed with the original samples (Fig. 6(b)), whereas mixing with samples from the standard DCGAN only slightly separated the samples (Fig. 6(c)). This indicates that the ResNet DCGAN proposed in this study can increase the distinguishability of the original sample categories.

B. CLASSIFICATION RESULT USING MIXED SAMPLES

Because the main objective of this study was to investigate whether mixing synthetic samples generated using GAN with original samples could improve the classifier, we fixed the classifier architecture to the ResNet BiLSTM architecture, which was shown to achieve good emotion classification results in our previous study [30], and compared the classification accuracy before and after data augmentation using synthetic samples.

This study employed the 80:20 validation technique, where 80% of the samples used for training were also used to train the DCGAN module. The synthetic samples generated by the DCGAN module were then mixed with the 80% training samples, but not with the validation samples (see Fig. 7). We also varied the proportion of the generated samples mixed with the original samples to determine whether different ratios would affect the classification results. To maintain a reasonable portion of the original samples, we set the maximum number of generated samples to be no more than double the number of original samples.

The method for calculating the proportion value was based on (14). For example, given 5600 original samples, we set aside 80% for training, which is 5600 × 0.8 = 4480 samples. We generated 8960 synthetic samples (double the original samples) and mixed a Portion, where Portion ∈ \{0.2, 0.4, 0.6, 0.8, 1\}, of these synthetic samples with the original samples. Therefore, for the case where we mix 4480 original samples with 8960 × 0.2 = 1792 synthetic samples, the proportion of generated samples would be 28.5%.

\[
\text{proportion of generated samples} = \frac{\# \text{ generated} \times \text{Portion}}{\# \text{ original} \times (0.8) + \# \text{ generated} \times \text{Portion}} \tag{14}
\]

Multiclass accuracy was employed as the comparison metric for both two-class and four-class classifications, calculated in the following manner:

\[
\text{multiclass accuracy} = \frac{\text{correctly classified}}{\text{sample size}} \tag{15}
\]

1) TWO-CLASS CLASSIFICATION RESULT

Table 6 outlines the classification accuracy results for the two-class emotion classification (positive and negative) using the ResNet BiLSTM architecture with datasets mixed with different proportions of synthetic samples. The highest accuracy was achieved with synthetic samples generated by the ResNet DCGAN architecture using a mixing proportion of 44.44%, and the accuracy increased from 87.47% to 90.34%. From Fig. 8 the trend of the classification accuracy for different mixing proportions shows an inverted U-shape for both samples generated by the standard DCGAN or ResNet DCGAN. This trend can be attributed to overfitting, where the classifier begins to overfit the set of generated samples, and thus no longer generalizes as well as it should. We can conclude that the addition of synthetic samples generated using the proposed DCGAN method improved the performance of the classifier.

2) FOUR-CLASS CLASSIFICATION RESULT

Table 7 and Fig. 9 present the results of the four-class classification (neutral, angry, happy, and sad). The highest
classification accuracy was again achieved at a mixing proportion of 44.44%, with ResNet DCGAN improving the accuracy from 82.06% to 86.32%, whereas the standard DCGAN achieved 85.96%.

From both the two-class and four-class classification tasks, we can conclude that the two proposed DCGAN modules generate samples that improve the classification accuracy using the ResNet BiLSTM model. The ResNet DCGAN outperformed the standard DCGAN in generating more distinguishable samples for emotion classification.

C. ABLATION STUDY
To study the individual impact of each module (ResNet, BiLSTM, and DCGAN) on the classification accuracy, an ablation study was performed using the 80:20 validation method, and the results are summarized in Table 8. It is evident that the integration of ResNet and BiLSTM outperforms individual classifiers, including the baseline model VGG-16 [46], and using ResNet or BiLSTM alone. The use of DCGAN for data augmentation further increased the accuracy of both two-class and four-class emotion classifications. Each individual module contributed to better accuracy, whereas the joint effect provided the best classification result, substantiating the efficacy of the proposed architecture and DCGAN data augmentation. Comparatively, the ResNet DCGAN generates more distinguishable samples than the standard DCGAN, as depicted in Fig. 5 and Fig. 6, and contributes to a relatively higher accuracy.

D. COMPARISON WITH RELEVANT WORKS
Most previous studies employing PPG or heart-related signals performed two-class emotion recognition, namely positive and negative, in the valence dimension. Therefore, we compared these with the two-class recognition results in this study, as shown in Table 9.

The work of Lee et al. [27] is most closely related to our work because PPG is used as the physiological signal of choice. The DEAP dataset was used in this study. However, they used the 1D PPG signal directly as the input to the model, in contrast to our method that converts PPG signals into spectrogram images for classification. A 1D convolutional neural network (1D CNN) was used to extract features from the 1D PPG signal and yielded 75.3% classification accuracy in the valence dimension. Santamaria-Granados et al. [47] employed ECG and GSR signals from the AMIGOS dataset and used CNN with features extracted from the R peaks of ECG and skin conductance response (SCR) peaks of GSR, which yielded 71% recognition accuracy, also in the two-class valence dimension.

However, it should be noted that these studies used different datasets and validation methods, and DCGAN data augmentation was used in our case. Therefore, a direct comparison may not be as meaningful. Nevertheless, the performance of the proposed ResNet BiLSTM model with DCGAN data augmentation with 90.34% accuracy in two-class emotion recognition in the valence dimension using only PPG signals remained promising.
VI. CONCLUSION

This study proposed a data augmentation method based on the ResNet DCGAN model to generate synthetic PPG spectrograms to improve emotion classification using the ResNet BiLSTM classifier. We compared the ResNet DCGAN with the standard DCGAN to augment the tasks of two- and four-class emotion classification. The experimental results show that synthetic samples generated by the ResNet DCGAN improve emotion classification for two-class from 87.47% to 90.34% and for four-class from 82.06% to 86.32%, which are better than those achieved by the standard DCGAN.

Five different measures were employed to evaluate the quality of the generated samples: the IS, FID, GQI, LDA, and MD. The results show that the samples generated by the standard DCGAN have better values for the IS, FID, and GQI measures, indicating that their distributions are closer to the original samples than those generated by the ResNet DCGAN. However, by comparing the MD values and visualizing the class distribution after performing LDA dimension reduction, we observed that the ResNet DCGAN was capable of identifying discriminative features of different classes, which translated into synthetic samples that better captured the qualities of each class, thereby improving the distinguishability of the classes. This explained the improved accuracy achieved by the classifier when it was trained with the addition of these synthetic samples. By comparing different proportions of generated samples when mixed with the original samples, the experimental results showed that mixing at 44.44% to 54.54% achieved the best improvement in classification accuracy.

However, the fact that the classification accuracy saturated at some mixing proportions and began to decrease when more synthetic samples were added to the mixture suggests that mode-collapse may have occurred. The generator over-optimized to the discriminator so that it synthesized only a few modes of output that suited the discriminator. Different GAN models that address the mode-collapse problem such as WGAN [48], VEEGAN [49] and NuGAN [50] may be studied. Transferring the proposed framework onto a mobile device for emotion recognition in different use cases may also be explored.

REFERENCES

[1] A. Rosengren, S. Hawken, S. Ōunpuu, K. Sliwa, M. Zubaid, W. A. Almahmood, K. N. Blackett, C. Siththi-Amorn, H. Sato, and S. Yusuf, “Association of psychosocial risk factors with risk of acute myocardial infarction in 1119 cases and 15 648 controls from 52 countries (the INTERHEART study): Case-control study,” *Lancet*, vol. 364, no. 9438, pp. 953–962, Sep. 2004, doi: 10.1016/S0140-6736(04)17019-0.

[2] C. Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha, “A survey of AI-based facial emotion recognition: Features, ML & DL techniques, age-wise datasets and future directions,” *IEEE Access*, vol. 9, pp. 165806–165840, 2021, doi: 10.1109/ACCESS.2021.3131733.

[3] S. S. Panchal, A. Hiremath, and N. R. Toravi, “Automatic face emotion recognition,” in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Tirunelveli, India, Dec. 2018, pp. 402–405, doi: 10.1109/ICSSIT.2018.8748847.

[4] F. Goodarzi, F. Z. Rokhani, M. I. Saripan, and M. H. Marhaban, “Mixed emotions in multi view face emotion recognition,” in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Kuching, Malaysia, Sep. 2017, pp. 396–401, doi: 10.1109/ICSIPA.2017.8120643.

[5] H.-J. Lee and K.-S. Hong, “A study on emotion recognition method and its application using PPG image,” in *Proc. Int. Conf. Inf. Commun. Technol. Convoy. (ICTC)*, Jeju-Si, South Korea, Oct. 2017, pp. 370–372, doi: 10.1109/ICTC.2017.8191005.

[6] V. Upadhyay and D. Kotak, “A review on different facial feature extraction methods for face emotions recognition system,” in *Proc. 4th Int. Conf. Inventive Syst. Control (ICISC)*, Coimbatore, India, Jan. 2020, pp. 15–19, doi: 10.1109/ICISC7916.2020.9711712.

[7] C. Zhang and L. Xue, “Autoencoder with emotion embedding for speech emotion recognition,” *IEEE Access*, vol. 9, pp. 51231–51241, 2021, doi: 10.1109/ACCESS.2021.3069818.

[8] A. A. Abdelhamid, E.-S.-M. El-Kenawy, B. Alotaibi, G. M. Amer, M. Y. Abdelkader, A. Ibrahim, and M. M. Eid, “Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm,” *IEEE Access*, vol. 10, pp. 49265–49284, 2022, doi: 10.1109/ACCESS.2021.3131757.

[9] B. T. Atmaja, A. Sasou, and M. Akagi, “Speech emotion and naturalness recognitions with multitask and single-task learnings,” *IEEE Access*, vol. 10, pp. 72381–72387, 2022, doi: 10.1109/ACCESS.2022.3189481.

[10] E. Batbaatar, M. Li, and K. H. Ryu, “Semantic-emotion neural network for emotion recognition from text,” *IEEE Access*, vol. 7, pp. 111866–111878, 2019, doi: 10.1109/ACCESS.2019.2934529.

[11] F. Ren and T. She, “Utilizing external knowledge to enhance semantics in emotion detection in conversation,” *IEEE Access*, vol. 9, pp. 154947–154956, 2021, doi: 10.1109/ACCESS.2021.3128277.

[12] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, “Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network,” *IEEE Access*, vol. 8, pp. 61672–61686, 2020, doi: 10.1109/ACCESS.2020.2984368.

[13] F. Noroozi, C. A. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, “Survey on emotional body gesture recognition,” *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 505–523, Apr. 2021, doi: 10.1109/TAFFC.2018.2874986.

[14] A. V. Zaboleeva-Zotova, A. S. Bobkov, V. L. Rozaliev, and A. B. Petrovsky, “Automated identification of human emotions by gestures and poses,” in *Proc. Congr. Comput. Intel. 11th Brazilian Congr. Comput. Intel.*, Ipójucà, Brazil, Sep. 2013, pp. 300–303, doi: 10.1109/BRICS-CCI-CBIC.2013.57.

[15] Z. Shen, J. Cheng, X. Hu, and Q. Dong, “Emotion recognition based on multi-view body gestures,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 3317–3321, doi: 10.1109/ICIP.2019.8803460.

[16] D.-H. Kang and D.-H. Kim, “1D convolutional autoencoder-based PPG and GSR signals for real-time emotion classification,” *IEEE Access*, vol. 7, pp. 91332–91352, 2019, doi: 10.1109/ACCESS.2020.3031342.

[17] D. Nikolova, P. Petkova, A. Manolova, and P. Georgieva, “ECG-based emotion recognition: Overview of methods and applications,” presented at the Adv. Neural Netw. Appl. (ANNA), 2018, pp. 1–5.

[18] M. Liu, D. Fan, X. Zhang, and X. Gong, “Human emotion recognition based on galvanic skin response signal feature selection and SVM,” in *Proc. Int. Conf. Smart City Syst. Eng. (ICCSCE)*, Hunan, China, Nov. 2016, pp. 154947–154956, doi: 10.1109/ICCSCE.2016.0051.

[19] R. M. Mehmood, R. Du, and H. J. Lee, “Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors,” *IEEE Access*, vol. 5, pp. 14797–14806, 2017, doi: 10.1109/ACCESS.2017.2724555.

[20] C. Qing, R. Qiao, X. Xu, and Y. Cheng, “Interpretable emotion recognition using EEG signals,” *IEEE Access*, vol. 7, pp. 94160–94170, 2019, doi: 10.1109/ACCESS.2019.2928691.

[21] J. Chen, T. Ro, and Z. Zha, “Emotion recognition with audio, video, EEG, and EMG: A dataset and baseline approaches,” *IEEE Access*, vol. 10, pp. 13229–13342, 2022, doi: 10.1109/ACCESS.2022.3146729.

[22] P. J. Bota, C. Wang, A. L. N. Fred, and H. Placidio Da Silva, “A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals,” *IEEE Access*, vol. 7, pp. 140990–141020, 2019, doi: 10.1109/ACCESS.2019.2944001.
C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch,

Y. Ye, L. Wang, Y. Wu, Y. Chen, T. Zhang, and Z. Zhang, “GAN

K. He, X. Zhang, R. Sen, and J. Sun, “Deep residual learning for

M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,”

S. E. Eskimez, D. Dimitriadis, R. Gmyr, and K. Kumanati, “GAN-based

M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial

S. Siami-Namini, N. Tavakoli, and A. S. Namin, “The performance of

K. Simonyan and A. Zisserman, “Very deep convolutional networks for

E. Abdulhay, and N. Arunkumar, “Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS),”

IEEE Access, vol. 7, pp. 57–67, 2018, doi: 10.1109/ACCESS.2018.2883213.

O. Alpar and O. Krejcar, “Frequency and time localization in bio-

Shao-Wei Wang received the B.S. degree in computer and communication engineering from the National Kaohsiung University of Science and Technology, Kaoshuing, Taiwan, in 2018, and the M.S. degree in electrical engineering from the National Chung Cheng University (NCCU), Chiayi, Taiwan, in 2021. From 2020 to 2021, he was an Exchange Student at Hokkaido University, Hokkaido, Japan. His research interests include biomedical signal processing, and the application of artificial intelligence to biomedical problems. He is a Life Member of the Taiwanese Society of Biomedical Engineering.

Sung-Nien Yu (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1987 and 1991, respectively, and the Ph.D. degree in biomedical engineering from Case Western Reserve University, OH, USA, in 1996. He joined as a Faculty Member of the Department of Electrical Engineering, National Chung Cheng University (NCCU), Chiayi, Taiwan, in 1999, as an Assistant Professor. He was promoted to an Associate Professor, in August 2006. Since August 2010, he has been a Professor with the Department of Electrical Engineering, NCCU, where he has been working as the Department Chair, since August 2021. His research interests include biomedical signal processing, and the application of artificial intelligence to biomedical problems. He is a Life Member of the Taiwanese Society of Biomedical Engineering.

Yuh-Ping Chang received the B.S. degree in physics from the National University of Singapore, Singapore, in 2015, and the M.A. degree in religion and humanity from Tzu Chi University, Hualien, Taiwan, in 2018. He is currently pursuing the M.S. degree in electrical engineering with the National Chung Cheng University (NCCU), Chiayi, Taiwan.

From 2021 to 2022, he was a Research Assistant at NCCU. His research interest includes the development of explainable neuro-symbolic AI system with applications in medical diagnosis.